
МАСТЕРА ПСИХОЛОГИИ
Анна Анастази Сьюзан Урбина
ПСИХОЛОГИЧЕСКОЕ ТЕСТИРОВАНИЕ
7-е международное издание
ПИТЕР

A. Anastasi, S. Urbina

PSYCHOLOGICAL TESTING

PRENTICE HALL

А. Анастази, С. Урбина

ПСИХОЛОГИЧЕСКОЕ ТЕСТИРОВАНИЕ

7-е международное издание



Москва · Санкт-Петербург · Нижний Новгород · Воронеж
Ростов-на-Дону · Екатеринбург · Самара · Новосибирск
Киев · Харьков · Минск
2007

ББК 88.3в6
УДК 159.9.072
А64

Анастази А., Урбина С.

А64 Психологическое тестирование. — 7-е изд. — СПб.: Питер, 2007. — 688 с.: ил. — (Серия «Мастера психологии»).

ISBN 978-5-272-00106-1
5-272-00106-0

Классическая работа Анны Анастази «Психологическое тестирование» по праву считается «энциклопедией западной тестологии». При подготовке 7-го издания, выпущенного в США в 1997 году, текст книги был основательно переработан. Появилось несколько новых глав, написанных соавтором А. Анастази — С. Урбиной. Содержательные изменения отражают новейшие тенденции развития психологического тестирования, в том числе возрастающее влияние компьютеризации как фактора интеграции психологической науки в целом и методов тестирования в частности. В новом издании уделено значительное внимание компьютеризированному адаптивному тестированию, метаанализу, моделированию структурными уравнениями, использованию доверительных интервалов, кросс-культурному тестированию, применению факторного анализа в разработке тестов личности и способностей и другим широко используемым и быстро развивающимся понятиям и процедурам, которые будут оказывать влияние на психометрическую практику в XXI веке.

ББК 88.3в6
УДК 159.9.072

ISBN 0-02-303085-2 (англ.)
ISBN 978-5-272-00106-1

СОДЕРЖАНИЕ

ПРЕДИСЛОВИЕ К РУССКОМУ ИЗДАНИЮ

В 1982 г. издательство «Педагогика» выпустило русский перевод книги Анны Анастази «Психологическое тестирование», которую редакторы перевода — К. М. Гуревич и В. И. Лубовский — по праву назвали «энциклопедией западной тестологии». Выход книги такого масштаба — всегда событие, а если учесть время и место — событие, как модно сейчас говорить, знаковое, поскольку ее появление было тогда воспринято как снятие негласного табу на широкое использование тестов в практической работе психологов, дефектологов, педагогов и других специалистов. Хотя со времени принятия печально известного постановления ЦК ВКП(б) о педологических извращениях в системе наркомпросов прошло более 45 лет, в начале 1980-х гг. его последствия были еще весьма ощутимы в советской психологии и педагогике. Так или иначе, книга Анастази стала для многих из нас не только источником знаний, но и тем долгожданным глотком свободы, который партия и правительство расчетливо давали сделать советской интеллигенции, дабы она не деградировала в изоляции от остального мира.

С тех пор прошло почти 20 лет. Многое изменилось в нашем обществе, однако потребность в книгах такого уровня, к счастью, осталась прежней, а если говорить о психологах, то, возможно, даже возросла вместе со значительным увеличением их числа. Издание 1982 г. с тиражом в 15 000 экземпляров стало библиографической редкостью почти сразу после выхода в свет. И хотя к настоящему времени выпущенный издательством «Педагогика» знаменитый двухтомник Анастази явно устарел, он по-прежнему пользуется большой популярностью у студентов, аспирантов и практических работников, связанных с тестированием. Мне не совсем понятно, почему наши — теперь уже не следующие директивам партии — издательства не воспользовались столь благоприятной маркетинговой ситуацией и не предприняли попыток выпустить перевод более свежего издания *Psychological Testing*, которая разошлась бы моментально. Возможно, потому что последнее, шестое, издание книги Анастази вышло в 1988 г. и могло показаться нашим издателям в середине 1990-х гг. не совсем «свежим». Возможно, были и другие причины, — книги, в которых встречаются формулы и графики, не пользуются любовью издателей. Как бы то ни было, это шестое издание оказалось на данный момент последним изданием «Психологического тестирования» Анны Анастази, ибо вышедшее в 1997 г. седьмое издание книги с тем же названием представляет собой в корне переработанный вариант, уже в соавторстве с Сюзаной Урбиной, и является, по существу, новой книгой. Именно этот вариант издательство «Питер» предложило мне для перевода.

Эта новая книга отличается от предыдущих изданий авторского учебника Анны Анастази в нескольких важных отношениях. Самые заметные перемены связаны с уменьшением объема. При незначительном изменении структуры глав, книга стала гораздо компактнее — теперь это один том, хотя и весьма солидный. Сокращению подверглись, в основном, подробности, касающиеся построения конкретных тестов и их психометрических характеристик. Однако тем, кто только начинает знакомиться

с такой сложнейшей областью деятельности, как психологическое тестирование, излишние подробности только мешают. К тому же конкретная информация о тестах, публикуемая в книгах учебного характера, даже для специалистов представляет скорее исторический интерес, поскольку устаревает с неимоверной скоростью. Для свежей информации существует периодика. Поэтому, если быть объективным, от сокращения объема книга только выиграла как учебник начального уровня.

С другой стороны, любое сокращение учебника влечет за собой снижение его самодостаточности. Современные учебники, особенно западные, встроены в систему информационного обеспечения через разветвленную систему ссылок на многочисленные источники, в которых подробно рассматриваются затрагиваемые в них вопросы. Книга Анастази и Урбины тоже построена в этом ключе и содержит обширную библиографию источников на английском языке, ссылки на которые даются практически в каждом абзаце текста. Для отечественных читателей это оборачивается двумя проблемами: получением доступа к таким источникам и необходимостью достаточно хорошо знать английский язык, чтобы быстро ознакомиться с их содержанием. Если последнюю проблему каждый человек решает самостоятельно, то решить первую проблему, даже с учетом развития Интернета, далеко не так просто. На мой взгляд, паллиативным решением могло бы быть создание собственного — минимального — информационного обеспечения для каждой заслуживающей того переводной книги. «Психологическое тестирование», несмотря на свой энциклопедический характер, относится к книгам типа «знаю что», и в этом ее достоинство. Но есть еще книги типа «знаю как», на которые, помимо нормативных документов, чаще всего и ссылаются А. Анастази и С. Урбина. К сожалению, именно таких книг не хватает нашим студентам, аспирантам и практикам. Разумеется, речь идет не о рецептурных поделках, а о серьезной литературе, написанной, однако, не для зрелых специалистов (такая литература у нас все же есть), а для тех, кто хочет стать таковым. Если говорить конкретно об издаваемой книге, то в качестве ее сопровождения хорошо было бы своевременно перевести ряд книг учебного характера по конструированию тестов, современным методам анализа заданий, метаанализу, многомерному шкалированию, факторному и кластерному анализу, методу моделирования структурными уравнениями, да и по отдельным типам тестов тоже. Это значительно бы повысило ее эффективность как учебника. Пока же в качестве такого сопровождения можно рекомендовать единственную книгу Пола Клайна «Справочное руководство по конструированию тестов» (1994), переведенную Е. П. Савченко под ред. Л. Ф. Бурлачука, — и то в принципе, поскольку она уже стала библиографической редкостью.

Изменения в содержании книги отражают основные тенденции развития психологического тестирования, которые авторы связывают с непрерывно возрастающим влиянием компьютеризации на эту область и с ее превращением в сферу политических и правовых интересов. Причем, как мне показалось, авторы придают компьютеризации статус фактора интеграции психологической науки в целом, и методов тестирования в частности, приводя в качестве примеров развитие психологической оценки, объединяющей два традиционно противопоставлявшихся подхода — психометрический и клинический; тенденцию к объединению нейропсихологических (косвенных) методов диагностики локальных поражений головного мозга с прямыми методами нейроинтроскопии; попытки учесть при разработке новых тестов неразрывное единство когнитивных и личностных переменных, и др. Все это действительно так, но, на мой взгляд, роль компьютеризации во всем этом гораздо скромнее. Интегративные

тенденции в психологии — результат ее собственного внутреннего развития. Психологи наконец-то стали понимать, что психоанализ и когнитивная психология — два *совместимых* аспекта описания сложнейшей природы человеческого поведения, а теория деятельности должна существовать не вместо всех научных школ, а *вместе* с другими подходами к объяснению человеческой психики. Компьютеризацию же лучше рассматривать как условие, облегчающее проявление интегративных тенденций, выстраданных в ходе короткой, но полной драматизма истории психологической науки.

Социальным, этическим и правовым аспектам тестирования в этом издании уделяется еще больше внимания, чем в прежних. Некоторые из затрагиваемых проблем, безусловно, специфичны для Америки и обусловлены содержанием конкретных законов. Тем не менее за всеми частностями скрывается ряд общих тенденций, крайне важных для развития психологического тестирования в нашей стране. Укажу лишь на одну, главную, на мой взгляд. Три составляющих ситуацию тестирования элемента — тестируемый, тестирующий и тест — по своему значению окончательно выстроились в указанном порядке. Из этого, казалось бы, тривиального факта вытекает множество отнюдь не тривиальных следствий методического, этического, социального и даже политического характера. С тестов постепенно снимаются обвинения во всех смертных грехах. Тесты — всего лишь инструменты в руках людей, и как любые другие инструменты — лопаты, пилы, топоры — могут быть хорошими и не очень, а иногда вообще бракованными. Для пользователей тестов все более очевидным становится факт, что все люди разные. Отсюда непременным условием подбора тестов, организации тестирования и, что особенно важно, интерпретации его результатов является учет истории развития индивидуума, особенностей его жизненного опыта и другой информации, релевантной целям тестирования. Взгляд на тестирование как экономящий время заменитель психологической оценки индивидуума уходит в прошлое. Все это резко повышает профессиональную, моральную и правовую ответственность тестирующего, распространяющуюся на весь процесс тестирования, от выбора подходящего для конкретных целей и конкретного человека теста до сообщения заключения по результатам теста получателю. В связи с этим повышаются и квалификационные требования к пользователям тестов. К слову сказать, просто купить профессиональный тест в Америке, пожалуй, сложнее, чем револьвер, поскольку в обществе давно осознали, насколько опасным в руках неопытных или безответственных людей может быть этот психологический инструмент. Вряд ли нужно убеждать читателей в остроте и актуальности подобных вопросов для сложившегося в нашей стране положения дел в области психологической практики.

Работая над переводом этой книги, я, естественно, пользовался русским изданием 1982 г., которое представляет собой перевод с четвертого издания *Psychological Testing*, вышедшего в 1976 г. Когда я сравнил оба оригинала — 1976 и 1997 гг., — то оказалось, что их текст, в среднем, совпадает примерно на 50% (естественно, в одних главах этот процент значительно меньше, в других — больше). Это вселяло оптимизм, сокращая работу вдвое. Однако, сравнив русский перевод издательства «Педагогика» с книгой, которую мне предстояло перевести, я обнаружил, как это ни покажется странным, гораздо меньше совпадений. Парадокс объясняется просто — временем. Этот перевод делался в конце 1970-х — начале 1980-х гг., и он просто устарел — как в отношении научного языка, так и в плане смысловых акцентов. К тому же текст глав, посвященных статистическим аспектам тестирования, содержал изрядное количество терминоло-

гических ошибок и смысловых неточностей, допущенных (по понятным причинам) переводчиками и пропущенных (по непонятным причинам) редакторами. Поэтому ничего не оставалось, как перевести всю книгу заново, сохраняя в совпадающих частях отдельные предложения и небольшие куски из старого перевода в тех случаях, когда они вписываются в современное прочтение текста.

Если говорить о трудностях перевода, то основная и, пожалуй, единственная трудность связана с лексически точным и кратким переводом названий тестов, нормативных документов, организаций и законов. В этой области нет устоявшихся образцов, зафиксированных в словарях, и потому возможны многочисленные варианты. Что касается названий тестов, то, как справедливо подчеркивают авторы этой книги, по ним нельзя судить о том, что измеряет тот или иной тест. Тем не менее большинство непрофессионалов судят о тестах как раз по их названию. Проблема усугубляется тем, что даже на языке оригинала названия тестов далеко не всегда точно соответствуют их содержанию и назначению, а при переводе вносятся дополнительные искажения. К примеру, вряд ли стоит называть тест, проверяющий понимание элементарных законов механики, изучаемых в средних классах школы, «тестом технических способностей», как это делается в русском издании 1982 г. В этом издании при переводе названий тестов я следовал, во-первых, принципу точности, и только во-вторых — принципу «красивости» названий товара (а то, что тесты — это товар, теперь хорошо известно и российским пользователям). В скобках после каждого названия теста, документа, организации или важного термина приведены соответствующее название или термин на языке оригинала. Это не только позволяет проверить работу переводчика, но и выполняет роль вспомогательного словаря для тех студентов и аспирантов, которые читают литературу по психологии на английском языке. Что касается математико-статистических терминов, то, в целом, они приведены в соответствие со стандартной терминологией в данной области.

Надеюсь, что эта книга послужит благородному делу преумножения знаний и повышению профессиональной культуры в области психологического тестирования, по меньшей мере, для нескольких поколений наших студентов, аспирантов и практических работников.

9 января 2001 г. *А. А. Алексеев*

ПРЕДИСЛОВИЕ

Памяти Джона Портера Фоули-младшего, внесшего существенный вклад в подготовку всех предыдущих изданий этой книги, посвящается.

Анна Анастази
Сюзанна Урбан

Девяностые годы свидетельствовали о неуклонном повышении и расширении интереса к психологическому тестированию, обнаружившегося в восьмидесятые. На это указывает как разработка новых тестов, часть которых отражает принципиально новые подходы, так и непрекращающиеся исследования существующих тестов наряду с систематическим пересмотром их более ранних версий. Главная цель, которую мы преследовали при отборе психодиагностического инструментария, заслуживающего упоминания или обсуждения на страницах этой книги, — раскрыть перед читателями многообразие измерительных инструментов, доступных в этой области на сегодняшний день, а также показать ряд тестов и методик, важных в историческом плане. Любая попытка дать исчерпывающее описание всей области психологического тестирования, или даже сколько-нибудь значительной ее части, потребовала бы книги иного объема.

Повышенное внимание уделяется людям, проходящим тестирование. Пользователи тестов побуждаются к поиску причин качества выполнения конкретного теста конкретным человеком в том, с какими событиями ему пришлось столкнуться в своей жизни и как он на них реагировал. Например, какие биографические сведения о данном человеке могли бы помочь понять его ответы на тест и повысить точность осуществляемого на основе полученных оценок прогнозирования последующего поведения — в школе, на работе и в других повседневных ситуациях? Из этого следует, что *пользователь теста* несет повышенную ответственность при выборе подходящих для конкретного человека тестов и методов проведения тестирования, равно как и при сообщении и использовании полученных результатов. Именно поэтому данный учебник задумывался, в основном, с целью обеспечить основу для правильного пользования тестами.

Эффективное использование тестов требует хотя бы элементарного знакомства с их конструированием. Такие знания необходимы для того, чтобы пользователь мог оценить различные тесты, выбрать среди них подходящие для определенных целей и конкретных обследуемых и правильно интерпретировать результаты тестирования. Хотя эта книга не адресована конкретно профессиональным разработчикам тестов, тем не менее, она содержит достаточно сведений о том, как создавать тесты, отвечающие потребностям пользователя.

В данном издании также даются простые объяснения некоторых широко используемых и быстро развивающихся понятий и процедур, которые, по всей вероятности, будут оказывать влияние на психометрическую практику в XXI в. Примерами таковых служат: компьютеризированное адаптивное тестирование, метаанализ, моделирование структурными уравнениями, использование доверительных интервалов вместо традиционной статистической значимости, кросс-культурное тестирование и все более широкое применение факторного анализа в разработке тестов личности и спо-

собностей. Применение различных моделей и техник факторного анализа в практике тестирования обеспечило получение норм, которые допускают интерпретацию оценок на разных уровнях специфичности или обобщенности, так что пользователь теста может выбрать тот уровень, который наиболее подходит для данного конкретного человека или конкретной ситуации.

В настоящее время в тестировании достаточно явно обнаруживаются две долгосрочные тенденции; вместо того чтобы посвятить им отдельные главы, мы обращаемся к их обсуждению на протяжении всей книги, всякий раз, когда рассматриваемый материал представляет для этого удобный случай. Первая тенденция — это постоянно возрастающее влияние компьютеризации на развитие, создание и проведение тестов, в добавление к твердо установившейся практике использования компьютеров для подсчета набранных баллов и последующей обработки результатов тестирования. Скорость технического прогресса столь велика, что он, по-видимому, опережает развитие существующих областей психологии. Однако технология оказывает мощное содействие психологии в ее продвижении на передовые позиции как в теории, так и в методах исследования. Например, в наше время происходит быстрое объединение и *«перекрестное оплодотворение» различных областей психологии, чему в немалой* степени способствует та легкость, с какой исследователи всего мира могут получать необходимую информацию, обрабатывать ее и обмениваться между собой полученными данными. Переосмысление когнитивных и личностных черт как взаимодействующих и неразделимых сторон индивидуума, который, в свою очередь, неотделим от его физического «Я», жизненных событий и среды, — один из самых ярких и многообещающих примеров этой тенденции к интеграции.

Вторая тенденция, весьма серьезно сказывающаяся на психологическом тестировании, отражает нарастающее вторжение политических и правовых интересов в эту область. Несмотря на то что эта тенденция сеет разногласия и несет в себе потенциальную опасность для развития тестологии, она все же имеет ряд позитивных последствий в виде побуждения к творчеству и повышенной бдительности в отношении планируемых и непредвиденных последствий использования тестов. Ссылки на ряд законов, оказавших влияние на практику тестирования, приведены на протяжении всего текста учебника, вместе с указанием их названий и года принятия; с их содержанием можно ознакомится по отчетам конгресса США и другим периодическим изданиям, которые можно найти в справочных отделах большинства библиотек.

Если на обложке первых шести изданий этого учебника стояло имя одного автора, то седьмое его издание подготовлено в соавторстве. Два автора вместе составляли план реорганизации глав и перечень охватываемых ими главных тем. Конкретная работа по пересмотру и переписыванию глав учебника была распределена следующим образом: Анастази — главы 1–7 и 10–12, Урбина — главы 8, 9 и 13–18. Кроме того, Урбина взяла на себя основные административные функции и ведение переписки. Однако, каждый из авторов знакомился с черновыми вариантами глав другого и вносил предложения, которые обычно принимались и вносились в окончательный текст книги.

Очевидно, что эта книга не могла быть написана, если бы авторы не имели доступа к результатам исследований и публикациям многих психологов из различных уголков США и других стран. Их имена встречаются на протяжении всей книги: при цитировании публикаций, при указании источников конкретных данных и в сводном перечне ссылок на использованную литературу. Внутри этой впечатляющей группы не-

сколько человек все же выделяются на общем фоне благодаря своей постоянной готовности к бескорыстному сотрудничеству и величине сделанного ими вклада в наше общее дело. Среди них мы должны в первую очередь упомянуть Дайану Браун (Dianne Brown) из научной дирекции Американской психологической ассоциации (*APA*), Аурелио Прифитеру (Aurelio Prifitera) и Джоан Ленке (Joanne Lenke) из Психологической корпорации (*Psychological Corporation*), Лоран Летандр (Lorin Letendre) из издательства *Consulting Psychologists Press*, Кэрол Уотсон (Carol Watson) из корпорации *NCS* (*National Computer Systems*), Дугласа Джексона (Douglas Jackson) из корпорации *SAS* (*Sigma Assessment Systems*), Элизабет Мак-Грэт (Elizabeth McGrath) и Джона Освальда (John Oswald) из издательства *Riverside Publishing Company*, а также Уэйна Камару (Wayne Camara) из Совета колледжей (*College Board*). Наконец, мы выражаем глубокую благодарность персоналу библиотек Университета Фордхама и Университета Северной Каролины за удовлетворение наших запросов, постоянно менявшихся в процессе работы над этой книгой.

А. Анастази
С. Урбина

Часть 1

ФУНКЦИИ И ИСТОКИ ПСИХОЛОГИЧЕСКОГО ТЕСТИРОВАНИЯ



1 ПРИРОДА И НАЗНАЧЕНИЕ ПСИХОЛОГИЧЕСКИХ ТЕСТОВ

Психологические тесты — это инструменты или, употребляя более широкий термин, орудия. Чтобы получить положительные результаты от применения тестов, мы должны учитывать этот важный факт. Любой инструмент может быть орудием, приносящим пользу или наносящим вред, — в зависимости от того, как его используют. Тестирование развивалось и продолжает развиваться нарастающими темпами, оказывая эффективное содействие в решении все более широкого круга вопросов в различных сферах повседневной жизни.[1] Однако его развитие сопровождалось нереалистичными ожиданиями и неправильным применением некоторых тестов. Пользователям нужно знать, как оценить тот или иной тест. Насколько подходит этот тест для той конкретной цели, ради достижения которой он используется? Какую информацию он может дать о человеке, который его выполняет? Как результаты этого теста можно включить в цепочку данных, приводящую к выбору линии действия? Именно такого рода вопросы мы ставили на первое место при подготовке этой книги. Наша книга ориентирована не на специалистов-тестологов, а на всех тех, кто изучает психологию. В настоящее время просто необходимо обладать определенной базой знаний о тестах, причем это касается не только тех, кто конструирует тесты или проводит тестирование, но всех и каждого, кто использует результаты тестов в качестве главного источника данных при принятии решений в отношении себя или других людей.

Области применения и разновидности тестов

Традиционно назначение психологических тестов состояло в том, чтобы измерять различия между людьми или между реакциями одного и того же человека в разных условиях. Одной из самых ранних проблем, побудивших к разработке психологических тестов, было выявление умственно отсталых. И до сегодняшнего дня обнаруже-

[1] Что касается ясных и убедительных иллюстраций потенциальных вкладов психологических тестов с примерами из реальной жизни, см. Dahlstrom (1993b).

ние интеллектуальных дефектов остается важной областью применения определенных видов психологических тестов. Родственная область клинического применения тестов включает обследование лиц с тяжелыми эмоциональными расстройствами и другими типами нарушения поведения. Помимо этого, мощный импульс первоначальному развитию тестов был задан стремлением удовлетворить нужды образования. Имеются в виду знаменитые тесты Бине, с которых, собственно говоря, и началось интеллектуальное тестирование. В настоящее время школы входят в число основных пользователей тестов. Распределение детей по способностям с целью максимально использовать возможности разных типов школьного обучения, выявление умственно отсталых, с одной стороны, и одаренных учеников — с другой, образовательное и профессиональное консультирование учащихся средних школ и студентов колледжей, отбор в профессиональные и другие специальные школы — вот лишь некоторые примеры использования тестов в образовании.

Отбор и распределение персонала на промышленных предприятиях — еще одна важная область применения психологического тестирования. От оператора на линии сборки или делопроизводителя до управленцев высшего звена вряд ли найдется работа, для которой тестирование не оказалось бы полезным при решении вопроса о найме, распределении обязанностей, переводе на новое место, повышении по службе или увольнении. Разумеется, эффективное применение тестов в большинстве таких ситуаций, особенно касающихся высококвалифицированной работы, возможно лишь тогда, когда тесты используют в качестве дополнения к специальному собеседованию с кандидатами, создающему условия — в виде биографического контекста — для правильной интерпретации тестовых показателей конкретного кандидата. Тем не менее тестирование составляет важную часть полной программы управления трудовыми ресурсами. Весьма близкое по целям применение психологического тестирования имеет место в вооруженных силах при отборе и распределении военнослужащих. По сравнению с эпохой Первой мировой войной, когда предпринимались отдельные попытки психологического обследования новобранцев, разнообразие и масштабы применения психологических тестов в армии во время Второй мировой войны значительно увеличились. Впоследствии исследованиями по разработке тестов были охвачены все рода войск.

Использование тестов в индивидуальном консультировании постепенно расширилось от узконаправленных советов относительно учебных и профессиональных планов до рассмотрения всех аспектов жизни человека. Эмоциональное благополучие и эффективные межличностные отношения все более отчетливо выделяются в качестве целей консультирования. Также отмечается усиливающаяся тенденция к использованию тестов для улучшения самопонимания и личностного роста. В рамках такого применения тестов их показатели составляют часть информации, на основе которой человек принимает решения относительно себя самого и своей жизни.

Совершенно очевидно, что психологические тесты в настоящее время применяют при решении широкого круга практических проблем. Однако не следует забывать и о том, что такие тесты — важное средство фундаментальных исследований. К примеру, почти все проблемы в дифференциальной психологии требуют обращения к методикам тестирования как средству сбора данных. В качестве иллюстраций можно указать на исследования природы, характера и степени индивидуальных различий, структуры психологических черт, измерение групповых различий и выявление биологических и культурных факторов, связанных с различиями в поведении. Во всех таких областях

исследования, как, впрочем, и во многих других, точное измерение индивидуальных различий, ставшее возможным благодаря правильно построенным тестам, является необходимым условием работы. Кроме того, психологические тесты служат стандартизованными инструментами исследования таких разнообразных проблем, как возрастные изменения в развитии человека на протяжении всей его жизни, относительная эффективность разных методов обучения, результативность психотерапии, воздействие социальных программ и влияние средовых переменных на человеческую деятельность.

Столь разнообразные по своему назначению виды тестов различаются и по другим важным характеристикам. Прежде всего, они разделяются по способу проведения тестирования: индивидуальному (проводимому квалифицированным специалистом), групповому или компьютерному. Далее, тесты различаются по тем аспектам поведения, для измерения которых они предназначены. Некоторые из них нацелены на оценку когнитивных особенностей, или способностей, которые могут варьировать от общих способностей, таких как готовность извлекать пользу из учебной работы в колледже, до высоко специфичных сенсомоторных умений, необходимых для выполнения простой ручной операции. Другие тесты обеспечивают измерение аффективных переменных, или личности, включая эмоциональные и мотивационные характеристики, межличностное поведение, интересы, аттитюды и ценности.

При столь очевидном разнообразии природы и назначения психологических тестов, есть ли у них какие-то общие отличительные признаки? Чем психологические тесты отличаются от других методов сбора информации о людях? Ответ следует искать в некоторых принципиальных особенностях конструирования и применения тестов. Рассмотрению этих особенностей и посвящен следующий раздел.

Что такое психологический тест?

Выборочный анализ поведения. Психологический тест, в сущности, представляет собой объективное и стандартизованное измерение образцов (или проб) поведения. Психологические тесты, подобно наблюдениям или тестам в любых других науках, проводятся на малой *выборке* тщательно отобранных образцов поведения индивидуума. В этом отношении психолог идет почти тем же путем, что и биохимик, делающий свои заключения о составе крови пациента или качестве питьевой воды в микрорайоне на основе анализа одной или нескольких взятых им проб. Если психолог хочет проверить словарный запас ребенка, умение служащего выполнять арифметические вычисления или зрительно-двигательную координацию пилота, он предъявляет им репрезентативные наборы слов, арифметических задач или же тесты двигательных способностей и оценивает их реакции. Насколько адекватен тест изучаемому аспекту поведения, зависит, очевидно, от количества и характера заданий, образующих стимульный набор (или выборку заданий) данного теста. Так, арифметический тест, состоящий из 5 задач или включающий лишь вопросы на умножение, вряд ли может дать достаточно верное представление о счетных навыках взрослого человека, а словарный тест, составленный целиком из терминов игры в бейсбол, едва ли обеспечит надежную оценку полного словарного запаса ребенка.

Диагностическая, или *предсказательная, ценность* психологического теста зависит от того, насколько он может служить индикатором относительно широкой и важ-

ной области поведения. Измерение образцов поведения, непосредственно охватываемых данным тестом, очень редко оказывается, если вообще оказывается, целью психологического тестирования. Знание ребенком какого-то списка из 50 слов, так же как и выполнение конкретной серии из 20 арифметических задач, сами по себе не представляют большого интереса. Но если можно продемонстрировать близкое соответствие между знанием ребенком данного списка слов и его общим словарным запасом или же между показателем, полученным при решении арифметических задач претендентом на должность клерка, и качеством выполнения им счетных операций на работе, то используемые тесты отвечают своему назначению.

В этой связи следует отметить, что задания теста не обязательно должны иметь близкое сходство с поведением, для предсказания которого тест предназначен. Здесь важно только, чтобы между ними и поведением наблюдалось эмпирическое соответствие. Степень сходства между тестируемыми образцами поведения и прогнозируемым поведением достаточно широко варьирует. На одном полюсе континуума «сходство — различие» тест может полностью совпадать с какой-то частью предсказываемого поведения. В качестве примеров можно было бы привести словарный тест по иностранному языку, проверяющий знание учащимися 20 из 50 вновь выученных слов, или тест на знание правил дорожного движения для получения водительских прав. Однако задания тестов профессиональной пригодности, применяемых перед началом обучения специальности, уже меньше похожи на те, которые приходится выполнять на настоящей работе. На другом полюсе находятся проективные личностные тесты, такие как тест чернильных пятен Роршаха, в котором делается попытка на основе ассоциаций, возникающих у обследуемого человека при разглядывании чернильных пятен, предсказать, как он будет реагировать на других людей, эмоционально окрашенные раздражители и прочие сложные ситуации повседневной жизни. Несмотря на внешние различия, все эти тесты состоят из выборочных проб поведения индивидуума. И ценность каждого должна доказываться эмпирически устанавливаемым соответствием между характеристиками деятельности человека в ситуации тестирования и в других ситуациях.

Используемые в этой связи термины «диагноз» и «прогноз» являются довольно слабым дифференциальным признаком. Обычно прогноз ассоциируется с оцениванием во временнóй перспективе, — например, будущее выполнение индивидом какой-либо деятельности предсказывается исходя из результатов выполнения им теста в настоящее время. Вместе с тем, в широком смысле, даже диагноз наличных состояний, таких как умственная отсталость или эмоциональные расстройства, содержит предсказание того, как будет вести себя человек с тем или иным диагнозом в ситуациях, отличных от тестовых. В логическом отношении проще рассматривать все тесты как выборочное измерение поведения, на основе которого можно предсказать поведение в других случаях. Поэтому имеет смысл рассматривать разные виды тестов как вариации этой основной схемы.

Еще один момент, который следовало бы обсудить в самом начале, связан с понятием *способность (capacity)*. Вполне возможно создать тест, например, для предсказания успешности овладения французским языком еще до того, как конкретный человек приступит к его изучению. Такой тест был бы связан с выборочным анализом типов поведения, необходимых для освоения нового языка, но сам по себе не предполагал бы знания французского. Тогда можно было бы сказать, что этот тест измеряет «способность» или «потенциальные возможности» индивидуума к овладению французским

языком. Однако такие термины по отношению к психологическим тестам следует использовать с осторожностью. Только в том смысле, что выборка образцов настоящего поведения может быть использована как индикатор другого, будущего поведения, мы вправе говорить об измерении «способности» данным тестом. Ни один психологический тест не может измерить ничего, кроме поведения. Будет ли такое поведение эффективным показателем другого поведения, определяется только эмпирическим путем.

Стандартизация. Напомним, что мы начали с определения психологического теста как стандартизованного измерения. Стандартизация подразумевает *единообразие процедуры* проведения и оценки выполнения теста. Если мы хотим, чтобы показатели, полученные разными людьми, были сравнимыми, условия тестирования должны быть одинаковыми для всех. Такое требование — всего лишь конкретное применение принципа контролируемости условий любого научного наблюдения. В тестовой ситуации единственной независимой переменной часто оказывается сам обследуемый.

Чтобы обеспечить единообразие условий тестирования, создатель теста дает подробные указания по проведению каждого вновь разработанного теста. Формулирование таких указаний — важная часть стандартизации нового теста. Она включает точные указания относительно используемого стимульного материала, временных ограничений, устных инструкций испытуемому, пробных образцов заданий, допустимых ответов на вопросы обследуемого и других тонкостей проведения теста. На выполнение некоторых тестов может влиять множество других, не столь очевидных факторов. Так, зачитывая вслух инструкцию или задания, следует принимать в расчет скорость речи, тон голоса, интонацию, паузы и выражение лица. В тесте на обнаружение нелепостей, например, правильный ответ может быть невольно подсказан улыбкой или паузой после произнесения критического слова. Значение стандартизованной процедуры тестирования с точки зрения проводящего обследование специалиста будет обсуждаться в этой главе позднее, в связи с проблемами проведения теста.

Другой важный этап в стандартизации теста — установление *норм*. Психологические тесты не имеют заранее определенных стандартов их успешного или неуспешного выполнения; критерии выполнения каждого теста устанавливаются на основе эмпирических данных. В большинстве случаев тестовый показатель индивидуума интерпретируется на основе сравнения с оценками, полученными по данному тесту другими людьми. Как следует из самого этого термина, норма — это обычный, или средний, уровень выполнения. Поэтому, если нормальные 8-летние дети правильно решают 12 задач из 50 в тесте на типичное арифметическое рассуждение, значит, норма для 8-летнего ребенка по этому тесту соответствует 12 (очкам, баллам или каким-то другим произвольным «единицам» измерения). Показатели такого рода принято называть первичными оценками (или «сырыми» баллами). Они могут выражаться числом правильно решенных заданий, временем, необходимым для их выполнения, числом ошибок и другими объективными мерами, соответствующими содержанию теста. Такая первичная оценка ни о чем не говорит до тех пор, пока не получит выражение в единицах подходящих интерпретационных данных.

В процессе стандартизации теста его проводят на большой репрезентативной выборке лиц определенного типа, для работы с которыми он предназначен. Эта группа, называемая выборкой стандартизации, как раз и служит для установления норм. Такие нормы показывают не только средний уровень выполнения теста, но и относительную частоту различных по степени отклонений от среднего уровня в обе стороны, что

позволяет количественно оценивать величину превышения или отставания от среднего. Конкретные формы выражения таких норм рассматриваются в главе 3. Любая из этих форм позволяет охарактеризовать положение индивидуума относительно нормативной выборки или выборки стандартизации.

Следует попутно отметить, что нормы для личностных тестов устанавливаются в сущности таким же образом, как и для тестов способностей. Норма по личностному тесту совсем не обязательно соответствует наиболее желательному или «идеальному» варианту его выполнения, равно как и норма по тесту способностей практически не бывает представлена абсолютным показателем, выставляемым за безошибочное выполнение всех тестовых заданий. Для тестов обоих типов норма соответствует результатам их выполнения типичными, или средними, людьми. Например, в тестах, измеряющих «доминирование — подчинение», норма приходится на среднюю точку, отображающую степень доминирования или подчинения, проявляемую средним человеком. Подобным же образом в опроснике эмоционального приспособления (*emotional adjustment inventory*) норма обычно не соответствует полному отсутствию неблагоприятных или дезадаптивных реакций. Некоторое количество таких реакций свойственно большинству «нормальных» людей, входящих в выборку стандартизации, и потому норма должна отображать это количество реакций, свидетельствующих об отсутствии у большинства людей абсолютного контроля над своими эмоциями.

Объективное измерение трудности. Обращаясь к определению психологического теста, с которого началось его обсуждение, напомним, что тест был охарактеризован не только как стандартизованное, но и как объективное измерение. В каких конкретных отношениях такие тесты являются объективными? Некоторые аспекты объективности психологических тестов уже были затронуты при обсуждении стандартизации. В связи с этим отмечалось, что процедуры тестирования, вычисления первичных оценок по тесту и их интерпретации объективны в той мере, насколько они независимы от субъективных суждений специалиста, проводящего тестирование. Теоретически у любого конкретного человека оценка по тесту должна быть одной и той же независимо от того, кто проводит с ним данный тест. На самом деле это не совсем так, поскольку полная стандартизация и абсолютная объективность практически недостижимы. Но по крайней мере стремление к такой объективности составляет одну из целей при конструировании теста, и нужно признать, что приемлемый уровень объективности достигнут в большинстве созданных тестов.

Есть и другие важные отношения, в которых психологические тесты с полным основанием можно охарактеризовать как объективные. Определение уровня трудности одного задания или теста в целом основывается на объективных эмпирических процедурах. Когда А. Бине и Т. Симон составляли свою первую (Binet & Simon, 1905) шкалу для измерения интеллекта, они расположили входящие в нее 30 заданий в порядке возрастания трудности. Уровень трудности определялся путем опробования этих заданий на 50 нормальных и нескольких умственно отсталых детях. Задания, с которыми справилось большинство детей, *ipso facto*[1], расценивались как самые легкие; задания же, с которыми справилось относительно малое число детей, считались более трудными. С помощью такой процедуры был установлен эмпирический порядок трудности всех заданий. Этот пример из истории тестирования служит прообразом объек-

[1] В силу самого факта (*лат.*). — *Примеч. науч. ред.*

тивного измерения уровня трудности, ставшего теперь общепринятой процедурой при создании психологических тестов.

Не только расположение, но и отбор заданий для включения их в тест может определяться исходя из доли входящих в пробную выборку лиц, которые справляются с каждым заданием. Так, если наблюдается скопление заданий на любом из концов шкалы (т. е. на полюсах легкости или трудности), от части таких заданий можно отказаться. Аналогично, если какие-то отрезки шкалы оказываются пустыми или представленными малым числом заданий, можно добавить новые, чтобы заполнить образовавшиеся пробелы. Более формальные аспекты анализа заданий будут рассмотрены в главе 7.

Надежность. Насколько хорош данный тест? Действительно ли он отвечает своему назначению? Эти вопросы могут выливаться, — и время от времени действительно выливаются, — в длительные бесплодные дискуссии. Субъективные мнения, необоснованные предчувствия и личные пристрастия могут приводить одних к переоценке возможностей конкретного теста, а у других вызывать его упорное неприятие. Единственный способ дать окончательный ответ на подобные вопросы — эмпирическая проверка. *Объективная оценка* психологических тестов предполагает в первую очередь определение их надежности и валидности в строго заданных ситуациях.

В психометрии термин «надежность» по существу означает согласованность. Надежность теста есть согласованность оценок у обследуемых лиц при их повторном тестировании тем же самым тестом или его эквивалентной формой. Если измерение *IQ* ребенка в понедельник дает коэффициент интеллекта, равный 110, а в пятницу, при повторном тестировании, равный 80, то очевидно, что ни к одной из этих оценок нельзя отнестись с доверием. Аналогично, если в наборе из 50 слов кто-то правильно определил 40, а в другом, считающимся эквивалентным, наборе — только 20, то ни одна из этих оценок не может рассматриваться в качестве надежного показателя уровня вербального понимания у данного человека. Разумеется, возможно, что в обоих примерах ошибочной является только одна из двух оценок, но это может показать лишь последующее тестирование; из приведенных данных следует только то, что обе оценки одновременно не могут быть правильными. Для более конкретного вывода (верна одна из оценок или неверны обе) требуется дополнительная информация.

Прежде чем давать разрешение на широкое использование психологического теста, необходимо провести тщательную объективную проверку его надежности. Различные типы надежности тестов и соответствующие методы ее измерения рассмотрены в главе 4. Надежность может проверяться путем сравнения результатов теста, получаемых при его проведении на одних и тех же людях в различные моменты времени, с использованием разных наборов заданий, при смене лиц, проводящих или оценивающих его выполнение, а также при варьировании любых других релевантных условий тестирования. Очень важно точно указывать тип надежности и способ ее определения, поскольку один и тот же тест может изменяться при этом в различных аспектах. Кроме того, следует сообщать сведения о величине и характере выборки, на которой проверялась надежность теста. Такая информация дает возможность пользователям теста предсказывать, будет ли данный тест столь же надежен для той группы, в которой они собираются его применить, или же им следует ожидать снижения (повышения) его фактической надежности по сравнению с номинальной.

Валидность. Несомненно, самый важный вопрос относительно всякого психологического теста касается его валидности: действительно ли данный тест измеряет то, для

измерения чего он предназначен, и в какой степени? Валидность предусматривает прямую проверку того, насколько хорошо тест выполняет свою функцию. Для определения валидности обычно требуются независимые, внешние *критерии* всего того, что тест должен измерять. Например, если тест пригодности к обучению медицинским профессиям используется при отборе поступающих в медицинское училище, таким критерием, в общем, будет являться успешное окончание этого училища. В процессе валидизации данного теста его следовало бы провести на большой группе студентов в то время, когда они поступают в училище. Показателями результативности их обучения в медицинском училище могли бы служить получаемые каждым студентом отметки, характеристики преподавателей, успешное или неуспешное прохождение практики и завершение обучения. Такая сводная характеристика и служит критерием, с которым должны соотноситься исходные тестовые показатели студентов. Сильная корреляция, или высокий *коэффициент валидности,* означала бы, что студенты, имевшие высокие показатели по тесту, добивались в медицинском училище заметных успехов, а у имевших низкие показатели успехи были слабыми. Слабая корреляция указывала бы на плохое соответствие между тестовыми показателями и критериальной мерой и, следовательно, на низкую валидность теста. В данном случае коэффициент валидности дает нам возможность определить, насколько точно может быть предсказана на основе тестовых показателей эффективность (по заданному критерию) деятельности.

Валидность тестов, предназначенных для других целей, устанавливается сходным образом относительно подходящих для этого критериев. Например, для теста профессиональной пригодности валидность можно установить, основываясь на результативности работы группы персонала, нанятого на испытательный срок. Валидность батареи тестов, предназначенных для определения летных качеств, может быть установлена по результатам тренировочных полетов. Валидизация тестов, имеющих более широкое применение, производится относительно ряда независимо получаемых поведенческих индексов, и их валидность устанавливается только в ходе постепенного накопления данных из множества различных исследований.

Читатель, может быть, обратил внимание на кажущуюся парадоксальность понятия валидности теста. Если так необходимо наблюдать за людьми вне тестовой ситуации или как-то иначе получать объективные данные о том, что мы пытаемся предсказать с помощью теста, почему же не отказаться от самого теста? Ответ на этот вопрос нужно искать в различиях между группой, на которой производится валидизация теста, и группами, в которых данный тест будет со временем использоваться по его прямому назначению. Прежде чем предоставить тест пользователям, его валидность устанавливается на репрезентативной выборке испытуемых. Показатели этих испытуемых используются не по прямому назначению, а только в целях проверки создаваемого теста. Если валидность теста доказывается таким методом, его можно применять на других выборках уже при отсутствии критериальных мер.

И все же можно возразить, что нужно лишь подождать до тех пор, пока используемые в качестве критериальных мер результаты деятельности или поведения *любой* группы появятся сами собой и таким образом получить ту информацию, которую мы пытаемся предсказать с помощью тестов. Однако подобный образ действий в большинстве случаев потребовал бы неприемлемых затрат времени и энергии. Так, если бы мы захотели определить, кто из поступающих на работу справится с ней или кто из абитуриентов успешно закончит колледж, нам пришлось бы принять всех желающих

(или, в крайнем случае, сформировать из них случайную выборку) и дожидаться окончательных итогов! Тесты как раз и предназначены для того, чтобы свести к минимуму недопустимую расточительность такого образа действий — и его пагубное эмоциональное воздействие на людей. С помощью тестов можно оценить, с заданным пределом погрешности, актуальный уровень навыков, знаний и других релевантных характеристик индивидуума, составляющих предпосылку его будущей деятельности. И чем выше валидность и надежность теста, тем меньше будет относительная величина погрешности.

Конкретные проблемы, с которыми сталкиваются при определении валидности тестов разных типов, а также используемые при этом специальные критерии и статистические методы рассмотрены в главах 5 и 6. Однако один момент необходимо обсудить сейчас. Валидность показывает нам не только степень соответствия теста своему назначению. Фактически, она указывает нам, *что* измеряется тем или иным тестом. Анализируя данные валидизации, мы можем объективно определить, что же все-таки измеряет наш тест. Поэтому было бы правильнее определять валидность как меру нашей уверенности в том, что тест измеряет именно то, для измерения чего он предназначен. Несомненно, интерпретация тестовых показателей была бы более ясной и однозначной, если бы тесты всегда получали названия исходя из эмпирически установленных соотношений, по которым устанавливалась их валидность. Тенденцию к изменению в этом направлении можно увидеть в выборе таких названий, как «тест академической оценки» и «тест распределения персонала» вместо неопределенного — «тест интеллекта».

Почему необходим контроль за использованием психологических тестов?

«Могу ли я получить бланки теста Стэнфорд—Бине? Мой племянник на следующей неделе поступает в школу N., и мне бы хотелось немного поднатаскать его, чтобы он смог поступить».

«Чтобы усовершенствовать программу чтения в нашей школе, нам нужен культурно-свободный тест интеллекта, позволяющий измерять врожденный потенциал ребенка».

«Вчера вечером я ответил на вопросы интеллектуального теста, опубликованного в журнале, и получил *IQ*, равный 80, я думаю, что психологические тесты просто глупы».

«Моя соседка по комнате изучает психологию. Она дала мне личностный тест, по которому я оказалась невротичной. Я так расстроилась, что даже перестала ходить на занятия».

«В прошлом году вы давали нашим служащим с исследовательскими целями новый личностный тест. Нам бы теперь хотелось иметь их тестовые показатели для картотеки кадров».

Эти высказывания не выдуманы. Каждое взято из реальных случаев, перечисление которых легко может продолжить любой психолог. Эти высказывания иллюстрируют возможность неправильного использования или интерпретации психологических тестов, ведущих к представлению о тестах как о чем-то бесполезном или даже вредящем

обследуемому. Как любой научный метод или точный инструмент, психологические тесты обнаруживают свою эффективность только при правильном и умелом применении. В руках недобросовестного или неквалифицированного пользователя такие тесты могут причинить серьезный вред. Есть два главных аргумента в пользу контроля за использованием психологических тестов: а) гарантирование того, что тесты будут проводить только квалифицированные специалисты, а получаемые результаты будут правильно использованы, и б) предотвращение знакомства широкой публики с содержанием тестов, которое может существенно снизить их валидность.

Квалифицированный специалист по тестированию. Потребность в квалифицированном специалисте на каждом из трех основных этапов тестирования — при выборе теста, его проведении и подсчете баллов с последующей интерпретацией результатов — очевидна. Тесты нельзя выбирать, подобно косилкам для газонов, по каталогу, высланному почтой. Их невозможно оценить по названию, автору или каким-то другим идентификационным признакам. Разумеется, для оценки таких факторов, как цена, объемность и легкость транспортировки тестовых материалов, время тестирования, легкость и быстрота подсчета первичных оценок, никакой психологической подготовки не требуется; все эти сведения обычно приводятся в каталоге тестов, и их необходимо учитывать при составлении программы тестирования. Однако, для того чтобы тест выполнил свои функции, важно оценить такие его технические характеристики, как валидность, надежность, уровень трудности и нормы. Только так пользователи могут определить, насколько тот или иной тест пригоден для решения их специфических задач и насколько он подходит для той категории лиц, которую они планируют обследовать с его помощью.

Несколько раньше в этой главе, при предварительном обсуждении стандартизации теста, уже указывалось на важность должной подготовки специалиста, проводящего тестирование. Если мы хотим, чтобы результаты, получаемые при проведении одного и того же теста разными специалистами, были сопоставимы, или чтобы можно было оценить тестовый показатель конкретного человека исходя из опубликованных норм, требуется полное понимание необходимости точно следовать инструкциям, равно как и доскональное знание стандартных процедур. Не менее важен и тщательный контроль условий тестирования. Аналогично этому, неправильный или неточный подсчет «сырых» баллов может полностью обесценить тестовый показатель. При отсутствии надлежащих контрольных процедур ошибки при подсчете «сырых» баллов встречаются намного чаще, чем, по-видимому, принято думать.

Правильная интерпретация тестовых показателей требует всестороннего понимания самого теста, особенностей обследуемого человека и условий тестирования. Что именно измеряется — можно объективно определить, только соотнося тест со специфическими процедурами, на основе которых была устаиовлена его валидность. Столь же необходима и информация о надежности, особенностях группы, на которой устанавливались нормы, и т. п. Существенными для интерпретации любых показателей теста являются биографические сведения о проходящем тестирование человеке. За одинаковой оценкой у разных лиц могут стоять совершенно разные причины. Поэтому заключения, которые делают исходя из таких оценок, порой существенно различаются. Наконец, нельзя не упомянуть и такие особые факторы, влияющие на конкретный показатель, как необычные условия тестирования, временные эмоциональные или физические состояния тестируемого и его предыдущий опыт прохождения тестов.

Роль пользователя тестов. Важным результатом развития психологического тестирования в 1980-е и 1990-е гг. стало растущее признание ключевой роли пользователя тестов (Anastasi, 1990b). В этом контексте пользователь тестов — любой человек, который использует тестовые показатели как главный источник информации при принятии практических решений. Пользователь тестов может быть, а может и не быть специалистом по проведению тестов и обработке результатов тестирования. В качестве примеров пользователей можно назвать учителей, консультантов, чиновников системы образования, кадровиков на промышленных предприятиях и в государственных учреждениях. Львиная доля критики в адрес тестов направлена не на какие-то только им — как специфическим инструментам — присущие особенности, а на неправильное использование результатов тестирования недостаточно компетентными пользователями. Ряд нарушений в этой области вызван предпочтением сокращенных форм тестов, стремлением к быстрым ответам и простым шаблонным решениям реальных проблем. Спешка вследствие перегруженности работой может поддерживать доверие к таким средствам достижения цели. И все же наиболее частой причиной неправильного использования тестов, вероятно, являются недостаточные или ошибочные знания пользователей в области тестирования (Eyde, Moreland, Robertson, Primoff, & Most, 1988; Moreland, Eyde, Robertson, Primoff, & Most, 1995; Tyler & Miller, 1986).

Специальные комитеты национальных профессиональных организаций, работающие совместно с издателями тестов, все больше внимания уделяют предупреждению неправильного использования тестов. Ярким примером тому служит проект, осуществляемый Рабочей группой по выработке квалификационных требований к пользователям тестов (*Test User Qualifications Working Group*), хорошо известной по очаровательному акрониму *TUQWoG* (Eyde et al., 1988). Главной целью *TUQWoG* было выработать опирающийся на широкий опыт набор необходимых квалификационных требований к пользователям различных видов тестов, с тем чтобы издатели тестов могли включить эти требования в свои формы для покупателей. В результате интенсивных общенациональных исследований в рамках проекта *TUQWoG* за пять лет была создана впечатляющая база данных. Некоторые издатели тестов уже начали использовать эти данные в своих квалификационных формах для покупателей. Позднее была образована вторая рабочая группа, целью которой стала разработка нормативных документов и учебиых материалов для пользователей тестов иа осиове базы данных *TUQWoG*. Ставшая известной под названием *TUTWoG* (измененный акроним *TUQWoG*, в котором *Q* заменена на *T*, первую букву слова *training* — обучение), эта группа в качестве своего первого продукта подготовила — с профилактическими целями — обзор наиболее распространенных случаев неправильного использования тестов (Eyde et al., 1993). Более поздняя сводка таких случаев дана в Moreland et al. (1995).

Закрытая и открытая информация о тестах. Ясно, что если бы кто-то заучил правильные ответы на тест для проверки цветовой слепоты, то с помощью такого теста уже не удалось бы оценить цветовое зрение этого человека. При таких обстоятельствах данный тест полностью утратил бы свою валидность. Очевидно, доступ к содержанию тестов следует ограничивать, чтобы предотвратить умышленные попытки исказить результаты тестирования. Тем не менее в других случаях влияние осведомленности может быть менее явным, а тест может утрачивать валидность в результате действий искренне заблуждающихся лиц. Например, школьный учитель может, из лучших побуждений, натренировать свой класс в решении задач, сходных с задания-

ми интеллектуального теста, с тем «чтобы дети были хорошо подготовлены к проведению теста». Такое отношение учителя продиктовано простым переносом обычной процедуры подготовки к школьным экзаменам на ситуацию тестирования. Однако применительно к тесту интеллекта такая специальная тренировка или натаскивание, вероятно, приведет к повышению тестовых показателей, не оказывая сколько-нибудь заметного влияния на более широкую область поведения, замеры которого пытаются произвести с помощью данного теста. При таких обстоятельствах валидность этого теста как прогностического или диагностического инструмента снижается.

Обеспечение защиты конкретного содержания тестов от несанкционированного доступа не должно мешать оперативному сообщению информации о тестах лицам, проходящим тестирование, заинтересованным специалистам и широкой публике. Распространение такой информации служит нескольким целям. Во-первых, она рассеивает мифы и раскрывает «тайны», связанные с тестированием, и тем самым содействует преодолению широко распространенных заблуждений в отношении назначения тестов и значения их показателей. Ряд публикаций, распространяемых некоторыми крупными, специализирующимися на выпуске тестов издательствами, отличаются ясным изложением материала и предназначены именно для этой цели. Во-вторых, часть распространяемой информации имеет отношение к специальным процедурам конструирования и оценивания конкретных тестов; эти материалы содержат релевантные данные о надежности, валидности и других психометрических характеристиках тестов. Такие сведения обычно включаются в специальные руководства, подготовленные для каждого профессионального теста, доступ к которым открыт всем заинтересованным лицам.

В-третьих, распространение информации о тестах преследует еще одну цель — ознакомить тех, кому предстоит пройти тестирование, с типичной обстановкой и процедурами проведения разных тестов, рассеять тревогу и создать все условия для того, чтобы каждый из этих людей мог в полной мере проявить себя при выполнении того или иного теста. С этой целью подготовлена серия разъяснительных буклетов, часть которых носит общий характер, тогда как другие касаются конкретных тестов, таких как Тест академической оценки Совета колледжей (*College Board's Scholastic Assessment Test*). Эти материалы обсуждаются в одном из последующих разделов данной главы. Наконец, в-четвертых, сообщение определенной информации служит крайне важной цели — обеспечению обратной связи прошедшим тестирование лицам, касающейся их собственных результатов по любому тесту, который с ними проводился. Психологи всегда уделяли большое внимание способам сообщения такой информации в разных контекстах, добиваясь ее максимальной полезности и содержательности. Соответствующие процедуры рассмотрены в главах 17 и 18.

Распространение информации о тестах и тестировании имеет исключительно важное значение. Здесь обнаруживаются как полезные, так и вредные тенденции. Примером последних могут служить излишне поспешные попытки законодательных органов ввести в этой сфере правительственный контроль на местном и федеральном уровнях (Bersoff, 1981, 1983; B. Lerner, 1980b). Местные законы, регулирующие раскрытие связанной с тестированием информации, были приняты и начали действовать в конце 1970-х гг. в штатах Калифорния и Нью-Йорк. Закон штата Нью-Йорк, более жесткий по сравнению с калифорнийским, требовал полного раскрытия вопросов тестов и ответов на них в крупномасштабных программах тестирования для приема в высшие учебные заведения.

Поскольку такое требование раскрытия информации делает необходимым подготовку новой формы каждого теста при очередном проведении тестирования, это может повлечь за собой любое из целого ряда неблагоприятных последствий. Среди них, помимо менее значимых, — сокращение количества наличных данных тестирования за год, повышение платы, взимаемой с абитуриентов за тестирование, и снижение контроля качества, наблюдаемое как при конструировании тестов, так и при уравнивании оценок по тестам, проводимым в разное время. Стоит также отметить, что лишь очень немногие из прошедших тестирование пользуются возможностью ознакомиться с содержанием тестов и ответами на них, предоставляемую им законом о раскрытии информации, и что результаты повторного тестирования по другой форме теста не улучшаются сколько-нибудь существенно от такого ознакомления (Stricker, 1984). Целей, которые послужили мотивом предложения законов о раскрытии связанной с тестированием информации, можно достичь более эффективным и безвредным способом, а именно путем интенсификации доступных механизмов сообщения информации о тестах.

Проведение тестирования

Главная причина применения тестирования заключается в возможности обобщения выборочных образцов поведения, наблюдаемых в тестовой ситуации, на поведение в других, нетестовых ситуациях. Тестовый показатель должен помочь нам предсказать, как пациент будет себя чувствовать и действовать за пределами клиники, как студент будет учиться в колледже, а поступающий на работу — справляться со своими обязанностями. Любые влияния, специфичные для тестовой ситуации, вносят вклад в дисперсию ошибок и снижают валидность теста. Вот почему так важно выявить все связанные с тестированием влияния, которые могут ограничивать или уменьшать возможность обобщения результатов теста.

Рассмотрению оптимальных методик проведения тестирования можно было бы посвятить целый том, но такой обзор выходит за рамки данной книги. Кроме того, полезнее познакомиться с такими методиками в конкретной обстановке, поскольку обычно ни один человек не имеет дело со всеми формами тестирования, — от обследования младенцев до клинического тестирования больных психозами или проведения программ массового тестирования военнослужащих. Поэтому в задачи этой книги входит главным образом рассмотрение общих принципов проведения тестирования, а не специальных вопросов их реализации в конкретных условиях. Прекрасный пример такой реализации можно найти у Sattler (1988, chap. 5), всесторонне рассматривающего индивидуальное оценивание детей.

Подготовка к проведению тестирования. Наиболее важным условием правильного проведения тестирования является предварительная подготовка. При тестировании не должны возникать непредвиденные обстоятельства. Поэтому нужно принять специальные меры для того, чтобы зараиее предупредить возникновение возможных случайностей. Только так можно обеспечить единообразие процедуры тестирования.

Предварительная подготовка к сеансу тестирования принимает множество форм. При проведении большинства индивидуальных тестов важно заучить наизусть словесную инструкцию. Даже в групповом тесте, в котором инструкция испытуемым

обычно зачитывается лицом, проводящим тестирование, предварительное ознакомление его с текстом предупреждает неправильное прочтение, запинание и позволяет вести себя более непринужденно и естественно во время сеанса. Еще одним важным предварительным шагом является подготовка тестовых материалов. В индивидуальном тестировании, особенно при проведении тестов действия (*performance test*)[1], такая подготовка включает размещение необходимых материалов с тем, чтобы свести к минимуму их поиски или неловкое обращение с ними. Как правило, материалы должны располагаться на столе вблизи места тестирования таким образом, чтобы они были легко доступны проводящему тест, но не отвлекали внимания обследуемого. При использовании сложной аппаратуры часто возникает необходимость в ее периодической проверке и калибровке. При проведении группового тестирования все тестовые бланки, листы для ответов, специальные карандаши и другие материалы заранее должны быть тщательно проверены, пересчитаны и разложены на рабочих местах испытуемых.

Подробное знакомство с процедурой проведения конкретного теста — еще одна важная форма подготовки к тестированию. Что касается индивидуального тестирования, такая подготовка обычно осуществляется в виде практического обучения проведению определенного теста под руководством опытного специалиста (супервизора). В зависимости от характера теста и типа обследуемых лиц для такого обучения может оказаться достаточным нескольких показов и практических занятий, а может потребоваться более чем годичное обучение. При групповом тестировании и особенно при проведении массовых обследований такая подготовка может включать предварительный инструктаж экзаменаторов (*examiners*) и наблюдателей (*proctors*) с тем, чтобы каждый хорошо представлял свои функции. Обычно экзаменаторы зачитывают инструкции, следят за временем выполнения и руководят действиями одной из групп. Кураторы выдают и собирают тестовые материалы, следят за тем, чтобы испытуемые выполняли инструкции, отвечают в разрешенных инструкцией пределах на их вопросы, не допускают с их стороны обмана.

Условия тестирования. Стандартизация затрагивает не только словесные инструкции, время выполнения заданий, материалы и другие аспекты самих тестов, но и обстановку тестирования. Определенное внимание нужно уделить выбору подходящего для тестирования помещения. Оно должно быть изолировано от чрезмерного шума и всего, что отвлекает внимание; в нем необходимо создать подходящее освещение, обеспечить вентиляцию, организовать удобные рабочие места для испытуемых. Следует также принять специальные меры, предотвращающие прерывание тестирования. Установка на двери специального предупреждающего знака эффективна лишь тогда, когда все знают, что такой знак запрещает входить в помещение при любых обстоятельствах. При тестировании больших групп бывает не лишне запереть двери или поставить около них помощников, не позволяющих войти опоздавшим.

Важно четко представлять себе возможную степень влияния условий тестирования на тестовые показатели. Даже кажущиеся незначительными аспекты тестовой ситуации могут заметно влиять на выполнение теста. Например, такой фактор, как использование парт или кресел с откидным столиком, повлиял на результаты группо-

---

[1] В отечественной литературе отсутствует устоявшееся название для этой разновидности тестов. Их также называют практическими тестами или невербальными тестами, хотя ни один из русскоязычных терминов, в том числе и «тесты действия», отражая их отдельные признаки, не передает полного значения термина *performance test*. — *Примеч. науч. ред.*

вого тестирования учащихся средних школ; в группах, сидевших за партами, они оказались выше (T. L. Kelley, 1943; Traxler & Hilkert, 1942). Имеются также доказательства того, что тип использованных бланков для ответов может влиять на тестовые показатели (F. O. Bell, Hoff, & Hoyt, 1964). Поскольку так сложилось, что агентства, подсчитывающие первичные оценки по тестам, и агентства, занимающиеся обработкой данных тестирования, работают независимо друг от друга и выпускают собственные бланки для ответов на задания теста, то бывают случаи, когда при проведении тестирования вместо бланков, применявшихся в процессе стандартизации теста, используются бланки ответов, приспособленные для машинной обработки. Без эмпирической проверки эквивалентность таких бланков не может считаться чем-то само собой разумеющимся. При тестировании детей до 5-го класса использование *любого* отдельного бланка для ответов может значительно снизить тестовые показатели (Cashen & Ramseyer, 1969; Ramseyer & Cashen, 1971). Для детей этого возраста, как правило, предпочтительней, чтобы они просто отмечали свои ответы в тестовой тетради.

Еще более существенные различия, причем на любом возрастном уровне, обнаруживаются при предъявлении одних и тех же тестов в бланковом и компьютерном вариантах. Влиянию этих различий в проведении тестирования на нормы, надежность и валидность в зависимости от характера теста и особенностей популяции тестируемых лиц уделялось большое внимание. Были составлены специальные методические руководства, облегчающие пользователям оценку сопоставимости тестовых показателей, полученных при этих двух вариантах проведения тестов.

Множество других, менее очевидных условий тестирования также могут влиять на выполнение тестов способностей и личностных тестов. От того, проводит ли тестирование совершенно незнакомый испытуемым человек или кто-то из тех, кого они уже знают, могут существенно измениться их результаты (Sacs, 1952; Tsudzuki, Hata, & Kuze, 1957). В другом исследовании было обнаружено, что манера поведения экзаменатора, который улыбался, кивал головой в знак согласия, делал замечания типа «хорошо» и «отлично», явно влияла на результаты тестирования (Wickes, 1956). В проективном тесте, где от испытуемого требовалось написать истории к предъявляемым картинкам, присутствие психолога-диагноста в комнате часто приводило к снижению эмоциональной окрашенности содержания этих историй (Bernstein, 1956). При проведении теста на умение печатать на машинке претенденты на рабочее место печатали значительно быстрее, если тестировались в одиночку, по сравнению с тестированием в группах из двух и более человек (Kirchner, 1966).

Можно было бы без труда умножить число таких примеров, но и приведенных достаточно, чтобы сделать три главных вывода. Во-первых, необходимо придерживаться стандартизованных процедур даже в мелочах. Обязанность создателя теста и издателя — добиться того, чтобы такие процедуры были полно и достаточно ясно описаны в руководстве к тесту. Во-вторых, следует регистрировать любые нестандартные условия тестирования, какими бы второстепенными они ни казались. В-третьих, при интерпретации результатов теста важно учитывать условия тестирования. При всестороннем обследовании личности в процессе индивидуального тестирования опытный диагност иногда отступает от стандартной процедуры проведения теста, с тем чтобы получить особо интересующую его дополнительную информацию. Поступив таким образом, он теряет право интерпретировать результаты теста на основе сопоставления с тестовыми нормами. В этом случае тестовые задания используются только для качественного исследования, а реакции испытуемого необходимо рассматри-

вать точно так же, как любые другие неформальные наблюдения за поведением или как данные интервью.

Начальный этап тестирования: раппорт и ориентирование испытуемого. В контексте проведения тестирования термин «раппорт» относится к попыткам проводящего тест специалиста вызвать у испытуемых интерес к тесту, добиться от них сотрудничества и содействовать тому, чтобы их реакции соответствовали целям теста. В соответствии с целью тестов способностей от испытуемых ожидают полного сосредоточения на предъявляемых задачах и приложения всех сил для того, чтобы хорошо их решить; цель личностных опросников предполагает искренние и честные ответы на вопросы, касающиеся повседневной жизни и обычного поведения; цели некоторых проективных тестов требуют полного отчета об ассоциациях, вызываемых тестовыми стимулами, без какого-либо их цензурирования или редактирования их содержания. Другие типы тестов могут требовать иных подходов. Но во всех случаях проводящий тестирование специалист старается побудить респондентов следовать инструкциям как можно добросовестнее.

Практическая подготовка специалистов по тестированию, помимо овладения методиками проведения различных тестов, предусматривает и обучение приемам установления раппорта. При установлении раппорта, так же как и при других процедурах тестирования, единообразие условий — существенный фактор получения сравнимых результатов. Если ребенку дают желанную награду за каждую правильно решенную тестовую задачу, его результаты нельзя сравнивать непосредственно с нормой или с результатами других детей, которых побуждали к решению лишь обычным словесным подбадриванием или похвалой. Любое отклонение от стандартных условий мотивирования в конкретном тесте следует отмечать и принимать во внимание при интерпретации результатов.

Хотя при индивидуальном тестировании может устанавливаться более полный раппорт, чем при групповом тестировании, в последнем случае все же стоит предпринять определенные шаги, с тем чтобы создать у испытуемых положительную мотивацию и уменьшить их тревогу. Специфические приемы установления раппорта варьируются в зависимости от характера теста, а также возраста и других характеристик тестируемых лиц. При тестировании дошкольников следует учитывать такие факторы, как боязнь незнакомых людей, легкую отвлекаемость и негативизм. Дружеская, веселая и мягкая манера поведения проводящего обследование специалиста помогает ребенку успокоиться. Пугливому, застенчивому малышу требуется больше времени для того, чтобы привыкнуть к новой обстановке. Поэтому лучше, если проводящий обследование не будет с самого начала слишком настойчивым, а подождет того момента, пока ребенок вступит с ним в контакт. Периоды тестирования должны быть непродолжительными, а тестовые задачи — разнообразными и интересными для ребенка. Тестирование должно проводиться в форме игры, и каждое предлагаемое ребенку задание должно возбуждать его любопытство. Процедура тестирования для этого возрастного уровня должна обладать достаточной гибкостью, позволяющей учитывать возможные отказы, утрату интереса и другие проявления негативизма.

Тестированию детей первых двух или трех классов начальной школы во многом свойственны те же трудности, что и тестированию дошкольников. Игровой подход по-прежнему остается наиболее эффективным способом возбуждения их интереса к тесту. Школьников постарше обычно можно мотивировать, обращаясь к свойственно-

му им духу соревновательности и желанию отличится при выполнении задания «учителя». Однако в тех случаях, когда тестируют детей, отстающих в обучении или выросших в иной культурной среде, не следует ожидать, что их стремление превзойти других по решению академических задач будет настолько же сильным, как и у детей из выборки стандартизации. Эта и другие проблемы, возникающие при тестировании лиц с несхожим жизненным опытом, рассмотрены в главах 9, 12 и 18.

С проблемами специфического взаимодействия мотивационных факторов можно столкнуться при тестировании лиц с эмоциональными нарушениями, заключенных или малолетних правонарушителей. Особенно в тех случаях, когда обследование проводится в официальной обстановке, эта категория лиц часто обнаруживает такие неблагоприятные аттитюды, как подозрительность, неуверенность, страх или циничное равнодушие. Особенности их прошлого опыта могут, вероятно, столь же неблагоприятно отражаться и на выполнении самого теста. Например, вследствие прежних неудач и срывов в школе у многих из них могло сложиться враждебное, сопровождаемое чувством собственной неполноценности отношение к школьным задачам, на которые так похожи задания теста. Опытный психолог-диагност предпринимает специальные усилия, чтобы в таких условиях наладить контакт с обследуемыми. Во всяком случае, он должен быть чуток к такого рода трудностям и принимать их во внимание при интерпретации результатов тестирования и объяснении качества выполнения теста.

При тестировании школьников или взрослых следует иметь в виду, что каждый тест представляет собой скрытую угрозу престижу индивидуума. Поэтому сначала испытуемых следует успокоить. Полезно, например, объяснить, что никто не ожидает от них выполнения, тем более абсолютно правильного, всех заданий. В противном случае, по мере перехода от простых заданий к более трудным или при невозможности закончить какой-то субтест в отведенное время, испытуемого может охватить быстро нарастающее чувство провала.

Желательно также по возможности устранить элемент неожиданности из ситуации тестирования, так как все неожиданное и неизвестное обычно вызывает тревогу. Многие групповые тесты снабжены предваряющими пояснениями, которые зачитывает группе лицо, проводящее тестирование. Еще лучше объявить о тестировании за несколько дней до его начала и дать каждому испытуемому отпечатанную брошюру, в которой объясняены цель и характер тестов, даны общие советы относительно их выполнения и приводится несколько примеров заданий. Такие разъяснительные буклеты и брошюры постоянно предоставляются в распоряжение участников многих программ массового тестирования, наподобие тех, что проводит Совет колледжей.

При тестировании взрослых возникают некоторые дополнительные проблемы. В отличие от ребенка младшего школьного возраста взрослый не обязательно будет стремиться решить задачу только потому, что она перед ним поставлена. Поэтому гораздо важнее убедить взрослого принять цель тестирования в качестве своей цели, хотя это справедливо уже по отношению к учащимся средних школ и колледжей. Сотрудничества испытуемых обычно можно достичь, убедив в том, что в их же собственных интересах получить по тесту валидный показатель, т. е. показатель, верно отражающий, а не преувеличивающий или преуменьшающий их способности. Большинство людей понимают, что неверное решение, принятое на основе недостоверных тестовых показателей, может привести к последующим неудачам, потере времени и разочарованию в себе. Такой подход не только побуждает проходящих тестирование лиц постараться проявить себя в тестах способностей, но также снижает процент лжи-

вых реакций и склоняет к искренним ответам в личностных опросниках, поскольку в этом случае респонденты сознают, что в противном случае они сами и останутся в проигрыше. Конечно же, не в интересах человека оказаться зачисленным на тот или иной курс обучения в университете, для усвоения которого у него отсутствуют необходимые знания и умения, так же как и быть принятым на работу, которую он не может выполнять или которая не соответствует его психическому складу.

Характеристики тестирующего и ситуационные переменные

Всесторонние обзоры влияния характеристик тестирующего и ситуационных переменных на тестовые показатели периодически публикуются (Lutey, & Copeland, 1982; Masling, 1960; S. B..Sarason, 1954; Sattler, 1970, 1988; Sattler, & Theye, 1967). Хотя ряд фактов такого влияния установлен для объективных групповых тестов, большинство данных было получено в отношении проективных методик и индивидуальных тестов интеллекта. Влияние побочных факторов, вероятно, сильнее сказывается на работе с неструктурированными и неясными стимулами, либо с трудными и новыми заданиями, чем на четко регламентированной и хорошо усвоенной деятельности. В общем, дети более восприимчивы к влияниям проводящего тестирование специалиста и ситуационным переменным, чем взрослые; при обследовании дошкольников роль диагноста оказывается решающей. Эмоционально неуравновешенные и неуверенные в себе люди, по-видимому, в любом возрасте более подвержены влиянию таких факторов по сравнению с людьми уравновешенными.

Как показали многочисленные исследования, при индивидуальном выполнении тестов интеллекта или проективных тестов на показатели могут влиять многие переменные, относящиеся к разряду личных качеств проводящего тестирование специалиста: его возраст, пол, раса, профессиональный или социоэкономический статус, уровень подготовки и опыт работы, особенности личности и внешний вид. Несмотря на обнаружение нескольких значимых связей, результаты подобных исследований часто оказываются неубедительными или обманчивыми, потому что их экспериментальный план не позволял контролировать или изолировать влияние различных характеристик тестирующего и тестируемого. Отсюда вполне возможно смешивание эффектов двух или более переменных.

Что касается влияния поведения тестирующего непосредственно перед проведением и во время проведения теста на результаты тестирования, здесь получены более ясные и убедительные данные. Например, проверочные исследования выявили значимые различия в показателях по тесту интеллекта в зависимости от того, какие отношения — «теплые» или «прохладные» — складывались между тестирующим и тестируемым, а также от того, держал ли себя тестирующий напряженно и отчужденно или, напротив, естественно и непринужденно (Exner, 1966; Masling, 1959). Кроме того, вполне возможны значимые взаимодействия между характеристиками тестирующего и тестируемого, в том смысле, что одни и те же качества тестирующего или его манера поведения могут по-разному влиять на разных испытуемых в зависимости от индивидуальных особенностей последних. Подобные взаимодействия могут происходить и с переменными, относящимися к задаче, такими как тип теста, цель тестирования и инструкции испытуемым. Дьер (Dyer, 1973) дополнил этот перечень другими

переменными, обратив внимание на возможное влияние расхождения в восприятии функций и целей тестирования проводящим тест и проходящим тестирование.

Еще одно возможное направление непреднамеренного влияния лиц, проводящих тестирование, на реакции тестируемого связано с их собственными ожиданиями. Это всего лишь особый случай самоосуществляемого пророчества (Harris & Rosenthal, 1985; R. Rosenthal, 1966; R. Rosenthal & Rosnow, 1969). Эксперимент с тестом Роршаха служит прекрасной иллюстрацией этого эффекта (Masling, 1965). Выразившим добровольное согласие 14 аспирантам предлагалось выступить в роли диагностов, причем 7 из них между прочим сообщали, что опытные диагносты выявляют больше реакций типа *H* и *Hd* (человеческие фигуры и их части), чем реакций типа *A* и *Ad* (фигуры животных и их части), а 7 другим говорилось обратное. При этих условиях две группы диагностов получили от обследованных ими лиц значимо различающиеся соотношения ответов *A* (*Ad*) и *H* (*Hd*). Эти различия возникли несмотря на то, что ни аспиранты в роли диагностов, ни сами обследуемые не сообщали о каких-либо попытках оказать на них влияние. Более того, магнитофонная запись сеансов тестирования не выявила никакого словесного воздействия со стороны диагностов. Их ожидания, по-видимому, находили свое выражение в едва уловимых изменениях позы и выражения *лица, на которые и реагировали обследуемые люди.*

Помимо проводящих тестирование лиц, на выполнение теста могут существенно влиять и другие аспекты тестовой ситуации. Новобранцы, например, часто подвергаются тестированию вскоре после поступления на службу, в период интенсивного приспособления к незнакомой и стрессовой ситуации. В одном исследовании, предназначенном установить влияние акклиматизации к такой ситуации на выполнение теста, 2724 новобранцам была предъявлена классификационная батарея (*Navy Classification Battery*) только на девятый день после их прибытия в Тренировочный центр ВМФ США (L. V. Gordon & Alf, 1960). Когда результаты этой группы сравнили с результатами, полученными 2180 новобранцами, которых протестировали, как было принято, на третий день пребывания, показатели группы обследованных на девятый день оказались значительно выше по всем субтестам батареи.

То, чем занимаются испытуемые непосредственно перед тестированием, также может влиять на выполнение теста, особенно если это вызывает волнение, беспокойство, усталость или другие отрицательно сказывающиеся на тестировании состояния. При исследовании учащихся 3-го и 4-го классов были получены данные, свидетельствующие о том, что *IQ*, оцениваемый по тесту «Нарисуй человека», зависит от того, чем занимались дети на уроке перед проведением тестирования (McCarthy, 1944). В одном случае ученики писали сочинение на тему «Самое лучшее, что когда-либо случалось со мной»; в другом, те же ученики снова писали сочинение, но уже на тему «Самое худшее, что когда-либо случалось со мной». Во втором случае, когда тест следовал за деятельностью, связанной, вероятно, с тягостным эмоциональным опытом, средний *IQ* был на 4–5 пунктов шкалы ниже, чем в первом случае. Эти данные получили подтверждение в более позднем исследовании, проведенном специально для определения влияния на результаты теста «Нарисуй человека» непосредственно предшествующего тестированию опыта (Reichenberg-Hackett, 1953). В данном исследовании дети, получившие удовлетворение после успешного решения интересной задачи-головоломки и поощренные игрушкой или конфетой, показали при тестировании лучшие результаты по сравнению с детьми, имевшими эмоционально нейтральный или менее положительный предшествующий опыт. Сходные данные были получены В. Е. Дэвисом (W. E. Davis,

1969а, 1969b) на студентах колледжа. Результаты теста на арифметическое рассуждение значимо снижались, когда перед его проведением студентам сообщали, что они плохо справились с тестом на вербальное понимание, чего не наблюдалось в контрольной группе, где тест на вербальное понимание не предъявлялся перед проверкой их арифметических навыков, как и в другой группе, которая в обычных условиях прошла стандартный тест на вербальное понимание.

Ряд исследований был посвящен изучению влияния обратной связи в отношении тестовых результатов на последующее выполнение теста индивидуумом. В тщательно спланированном исследовании семиклассников Бриджмен (Bridgeman, 1974) установил, что сообщение об «успехе» значительно улучшало выполнение сходного теста по сравнению с сообщением о «неудаче», хотя испытуемые в действительности выполнили первоначальный тест одинаково хорошо. Этот тип мотивационной обратной связи может действовать, главным образом, через те цели, которые испытуемые ставят себе при выполнении последующих заданий, и потому может рассматриваться как еще один пример самоосуществляемого пророчества. Однако такую неспецифическую мотивационную обратную связь не следует смешивать с корректирующей обратной связью, посредством которой индивидуум информируется о допущенных им конкретных упущениях и получает инструкции по их исправлению; в таких условиях обратная связь, по всей вероятности, должна улучшать выполнение теста испытуемыми, чьи показатели первоначально были низкими.

Примеры, приведенные в этом разделе, демонстрируют широкое разнообразие связанных с тестом и влияющих на тестовые показатели переменных. В большинстве правильно проводимых программ тестирования влияние таких переменных практически не ощутимо. Тем не менее квалифицированный специалист по тестированию должен быть всегда начеку, чтобы вовремя обнаружить их возможное действие и свести его к минимуму. В тех случаях, когда обстоятельства не позволяют контролировать некоторые условия тестирования, заключения по результатам выполнения теста следует сопровождать смягчающими оговорками.

Тестирование глазами тестируемых

Тестовая тревожность. Работы, посвященные изучению тестовой тревожности, относятся к числу самых первых исследований реакций испытуемых на ситуацию тестирования. Безусловно, ранний интерес к этому типу реакции был вызван ее заметностью и ее явно пагубным воздействием на результаты тестирования. Многие приемы, предназначенные для улучшения раппорта во время проведения теста, способствуют также снижению тестовой тревожности. Процедуры, служащие для рассеивания опасений, вызываемых неестественностью ситуации тестирования и таящимися в ней неожиданностями, успокаивающие и ободряющие испытуемого, конечно же, помогают снизить и его тревожность. Манера поведения проводящего тестирование и хорошая организация — без сбоев и помех — всего процесса служат той же цели.

Индивидуальные различия в тестовой тревожности изучали, в основном, на учащихся школ и студентах колледжей (Gaudry & Spielberger, 1974; Hagtvet & Johnsen, 1992; I. G. Sarason, 1980; Spielberger, 1972). У истоков большинства этих исследований стояли С. Б. Саразон и его коллеги по Иельскому университету (Sarason, Davidson, Lighthall, Waite, & Ruebush, 1960). Первым шагом явилось создание вопросника

для оценки аттитюдов тестируемого. Форма для детей содержала, например, такие вопросы:

> Сильно ли ты волнуешься перед тестированием?
> Когда учительница говорит, что она собирается проверить, как много вы выучили, начинает ли твое сердце биться быстрее?
> Во время выполнения теста думаешь ли ты о том, что у тебя не очень хорошо получается?

Самое интересное из обнаруженного исследователями представляет тот факт, что как показатели тестов школьных достижений, так и показатели интеллектуальных тестов имеют значимые отрицательные корреляции с тревожностью. Сходные корреляции были получены и на выборке студентов колледжей (I. G. Sarason, 1961). Лонгитюдные исследования также подтвердили существование обратной зависимости между изменениями в уровне тревожности и изменениями в выполнении тестов достижений и тестов интеллекта (K. T. Hill & S. B. Sarason, 1966; S. B. Sarason, K. T. Hill, & Zimbargo, 1964).

Конечно, такие данные ничего не говорят о направлении причинных связей. Возможно, учащиеся проявляют тревожность при тестировании из-за того, что плохо справляются с тестами и, таким образом, уже приобрели опыт неудач и разочарований в предыдущих ситуациях тестирования. В подтверждение такого объяснения можно привести данные, что внутри подгрупп с высокими показателями по тестам интеллекта обратная зависимость между уровнем тревожности и успешностью выполнения теста исчезает (Denny, 1966; Feldhusen & Klausmeier, 1962). С другой стороны, есть данные, свидетельствующие о том, что по крайней мере частично эта зависимость является результатом вредного влияния тревожности на выполнение теста. В одном исследовании (Waite, Sarason, Lighthall, & Davidson, 1958) высоко- и низкотревожным детям, уравненным по показателям теста интеллекта, давали повторные попытки в выполнении задания на научение. Несмотря на первоначально одинаковые успехи в выполнении этого теста научения, группа низкотревожных детей существенно улучшила свои результаты по сравнению с группой низкотревожных.

Некоторые исследователи сравнивали выполнение тестов в условиях, специально создаваемых для того, чтобы вызвать тревогу или, наоборот, снять напряжение тестируемых. Мандлер и Саразон (Mandler & Sarason, 1952), например, обнаружили, что инструкции с личной направленностью (*ego-involving*), в которых подчеркивается ожидание проводящего тест, что все из проходящих тестирование успеют закончить работу в отведенное время, благотворно сказываются на выполнении теста низкотревожными и неблагоприятно влияют на высокотревожных. Другие исследования обнаружили, кроме того, взаимодействие условий тестирования с такими индивидуальными особенностями тестируемых, как уровень тревожности и мотивация достижения (Lawrence, 1962; Paul & Eriksen, 1964). По всей видимости, связь между тревожностью и выполнением теста носит нелинейный характер, небольшая тревога сказывается благотворно на результативности, а сильная — пагубно. Для низкотревожных испытуемых благоприятны тестовые условия, вызывающие состояние некоторой тревоги, тогда как высокотревожные обычно лучше справляются с тестом в более спокойном состоянии.

Несомненно, что хронически высокий уровень тревожности пагубно сказывается на школьном обучении и интеллектуальном развитии. Тревога мешает как приобрете-

нию знаний, так и поиску информации в памяти (Hagtvet & Johnsen, 1992). Однако такое воздействие тревоги следует отличать от ограниченных тестовой ситуацией эффектов, которые мы обсуждаем, т. е. от того, в какой степени тестовая тревожность меняет качество выполнения теста, характерное для данного конкретного человека вне тестовой ситуации. Доказано, что в условиях конкурентного давления, испытываемого выпускниками средних школ при поступлении в колледж, тестовая тревожность существенно влияет на качество выполнения ими вступительных тестов. В тщательном и хорошо спланированном исследовании этой проблемы Френч (French, 1962) сравнил выполнение выпускниками средней школы теста, представляющего собой часть обычно проводимого Теста академических способностей (*SAT*), с выполнением его параллельной формы, когда тестирование проводилось в другое время и в менее напряженной обстановке. Инструкция в последнем случае специально подчеркивала, что тест дается только с научно-исследовательскими целями и показатели по нему не будут передаваться в колледжи. Оказалось, что в стандартной ситуации экзаменов результаты учащихся по этому тесту были ничуть не хуже результатов, полученных ими в более спокойном состоянии. Кроме того, при этих двух условиях не было обнаружено значимых различий в текущей валидности (*concurrent validity*) тестовых показателей относительно отметок по входящим школьную программу предметам. Данные ряда современных исследований также заставляют усомниться в расхожем представлении о патологически боящихся тестов учащихся, которые знают предмет, но буквально «коченеют» во время тестирования (см. Culler & Holahan, 1980). В частности, эти исследования показали, что учащиеся с высокими показателями по шкале тестовой тревожности получают, в среднем, более низкие текущие отметки по предметам и обладают менее развитыми учебными умениями, чем учащиеся с низкими показателями по этой шкале.

Исследования природы, проблем измерения и способов снижения тестовой тревожности продолжались с нарастающими темпами (I. G. Sarason, 1980; Spielberger, Anton, & Bedell, 1976; Spielberger, Gonzalez, & Fletcher, 1979; Spielberger, Gonzalez, Taylor, Algaze, & Anton, 1978; G. S. Tryon, 1980). Что касается природы тестовой тревожности, то здесь были выделены два важных компонента, именно: эмоциональность и озабоченность. Эмоциональная составляющая тестовой тревожности охватывает чувства и физиологические реакции, такие как напряжение и увеличение частоты сердечных сокращений. Озабоченность, или когнитивная составляющая, включает связанные с собой негативные мысли, такие как ожидание неудачи при выполнении теста и озабоченность последствиями провала. Эти мысли отвлекают внимание тестируемого от заданий теста и тем самым нарушают его выполнение. Оба компонента тестовой тревожности измеряются специально разработанными для этой цели опросниками. Несмотря на их широкое применение в исследованиях, до настоящего времени с такими опросниками можно было познакомиться только по сообщениям в научной литературе. Разработанный Спилбергером и его сотрудниками Вопросник тестовой тревожности (*TAI*) — единственный пример опубликованного теста такого рода; он кратко описан в главе 13 и включен в перечень опубликованных тестов (приложение А).

Немало исследований было посвящено разработке и оценке методов избавления от тестовой тревожности, которые вобрали в себя ряд методик поведенческой терапии (главе 17) для сокращения ее эмоционального компонента. В общем, их результаты были положительными, хотя и трудно отнести наблюдаемое улучшение на счет какой-то конкретной методики из-за методологических упущений в оценочных ис-

следованиях (G. S. Tryon, 1980). Фактически, эмоциональный компонент тестовой тревожности имеет тенденцию убывать с каждым последующим тестированием даже в контрольных группах без терапевтического вмешательства, не говоря уже о специальных контрольных группах, в которых проводилась правдоподобная псевдотерапия. Кроме того, сокращение эмоционального компонента почти или совсем не влияло на уровень выполнения тестов.

Повышение результативности выполнения тестов, а заодно и учебной работы, чаще наблюдается в тех случаях, когда воздействие оказывается на когнитивные реакции индивидуума в отношении самого себя. Выполненные на данный момент исследования свидетельствуют о том, что наилучшие результаты достигаются при использовании программ комбинированного воздействия, нацеленных не только на устранение излишних эмоций и чрезмерной озабоченности, но и на совершенствование учебных умений. Тестовая тревожность — комплексный феномен, вызываемый множеством разнородных причин, относительный вклад которых варьирует от человека к человеку. Чтобы быть эффективными, программы вмешательства должны приспосабливаться к нуждам конкретных людей. К тому же нужно отдавать себе отчет в том, что тестовая тревожность — это только одно проявление более общего комплекса условий, *снижающих эффективность человека как ученика.*

Комплексное исследование отношений тестируемых к тестированию. Хотя тестовая тревожность является заметным и важным аспектом поведения тестируемых, в нем есть еще немало других аспектов, изучение которых могло бы принести существенную пользу. Вышедшая в 1993 г. книга под редакцией Баруха Нево (Baruch Nevo) и Р. С. Ягера (R. S. Jäger) представляет собой широкомасштабную попытку собрать воедино доступную информацию о реакциях обследуемых на тестирование в сферах образования, промышленности, медицины и консультирования. Пятнадцать ее глав подготовлены ведущими специалистами в области изучения различных аспектов и приложений тестирования на основе доступных публикаций ученых разных стран по каждой теме, включая, разумеется, данные собственных исследований авторов. Результатом этого труда стала серьезная, основанная на широком базисе фактов попытка ответить на вопросы, которые до этого рассматривались, главным образом, в журнальных статьях или в политических и юридических источниках. Таким образом, эта книга служит средством коррекции накопившегося к настоящему времени изрядного количества предвзятых и противоречивых мнений о тестировании. Например, первая глава посвящена изложению результатов десяти профессионально проведенных опросов с целью выявления аттитюдов в отношении тестирования в выборках, представляющих самые разные категории населения. Эти результаты обнаруживают расхождения между взглядами широких кругов населения на спорные вопросы тестирования и некоторыми заявлениями ораторов, имеющих выход на широкую аудиторию, но выражающих скорее свою узкую позицию по данным вопросам.

Отдельные главы охватывают большой диапазон тем. Несколько глав посвящены разработке и использованию вопросников обратной связи и методов группового интервью для оценивания отношения к предъявляемым тестам и понимания того, что эти тесты измеряют, в различных группах тестируемых. В одной главе сравниваются мнения учащихся в отношении свободной формы контроля и классных тестов, составленных из заданий с множественным выбором, и результаты этого сравнения показывают явное предпочтение учащимися последнего варианта. Некоторые авторы рассматривают реакции претендентов на вакантные рабочие места в отношении честно-

сти тестирования и связанности тестов с характером предлагаемой работы. В нескольких главах предлагаются основанные на опыте авторов пути и способы усовершенствования проведения тестов, а также улучшения обстановки тестирования. В целом, составляющие эту книгу главы раскрывают перед нами многообещающую область исследований, предпринимаемых с целью отыскать решения ряда текущих социальных и практических проблем современного тестирования. Кроме того, эта книга служит улучшению взаимопонимания между пользователями тестов и тестируемыми.

Влияние практического обучения на выполнение тестов

При оценивании влияния тренировки или практики на тестовые показатели основной вопрос заключается в том, ограничивается ли улучшение конкретными заданиями, включенными в определенный тест, или же оно распространяется на более широкую область поведения, для оценки которого и предназначен данный тест. Ответ на этот вопрос содержится в различии между практическим обучением (*training*) и тренировкой (*coaching*). Очевидно, что любой полученный индивидуумом учебный опыт, независимо от того, носит он формальный или неформальный характер, приобретен в школе или вне ее, должен отразиться на выполнении им тестов, которые выборочно проверяют релевантные аспекты поведения. Такое широкое воздействие никак не снижает валидность теста, поскольку тестовый показатель дает точную картину текущего статуса индивидуума в отношении исследуемых способностей. Разумеется, обсуждаемое различие — это различие в степени. Воздействия невозможно классифицировать на узкие или широкие, поскольку они значительно варьируют по своим масштабам: от воздействий, влияющих на единственное применение единичного теста, к воздействиям, сказывающимся на выполнении всех заданий определенного типа, до воздействий, изменяющих выполнение индивидуумом подавляющего большинства операций. Однако, с точки зрения эффективного тестирования, можно ввести рабочий критерий для разграничения воздействий учебного опыта. Так, например, можно принять, что тестовый показатель становится невалидным только в тех случаях, когда конкретный опыт повышает его, не оказывая при этом заметного влияния на область поведения, для измерения которого предназначен данный тест.

Тренировка. Влияние тренировки на тестовые показатели исследовалось достаточно широко. Несколько ранних исследований было проведено английскими психологами, которых особенно интересовало воздействие практики и тренировки на тесты, применявшиеся при распределении 11-летних детей в средние школы разного типа (Yates et al., 1953–1954). Как и можно было ожидать, степень улучшения зависела от способностей и предшествовавшего тренировке образовательного опыта, характера теста, а также количества и типа тренировок. Дети с пробелами в образовании, по всей вероятности, извлекали больше пользы из специальной тренировки по сравнению с детьми, получившими хорошее образование и, следовательно, уже подготовленными к тому, чтобы хорошо выполнить тесты. Очевидно также, что чем выше сходство между содержанием теста и материалом тренировки, тем большего повышения тестовых показателей можно ожидать. С другой стороны, чем меньше обучение выхо-

дит за пределы содержания конкретного теста, тем менее вероятно распространение улучшения на деятельность, результаты которой используются в качестве критериальной меры валидности этого теста. Попутно следует отметить, что многие исследования влияния тренировки на выполнение тестов дают неоднозначные и неинтерпретируемые результаты из-за серьезных методологических изъянов (Anastasi, 1981a; Bond, 1989; Messick, 1980a), главный среди которых — неспособность найти нетренированную контрольную группу, которая действительно была бы сопоставимой с тренированной группой. Например, учащиеся, записавшиеся на платные подготовительные программы, представляют собой самосформировавшуюся выборку и, в целом, отличаются от учащихся контрольной группы по начальному уровню способности, мотивации и другим личным качествам, которые влияют на выполнение теста. Далее, в экспериментальных планах, предполагающих использование тестирования до и после тренировки, трудно обеспечить одинаковую мотивацию испытуемых выполнить тесты как можно лучше в обоих случаях; и это практически не удается сделать, когда одно обследование проводится во время регулярной, формальной проверки знаний учащихся, а другое — в неурочное время и в неформальной обстановке, ради практики или с исследовательскими целями.

Совет по вступительным экзаменам в колледжи США (*College Entrance Examination Board*) был обеспокоен расширением числа недобросовестно работающих коммерческих подготовительных курсов для абитуриентов. Чтобы прояснить этот вопрос, Совет колледжей провел ряд хорошо спланированных экспериментов для определения влияния обеспечиваемой этими курсами тренировки (или, точнее, натаскивания) на выполнение Теста академических способностей (*SAT*), а также подготовил обзор результатов, полученных в аналогичных исследованиях другими, независимыми специалистами (Donlon, 1984; Messick, 1980a, 1981; Messick & Jungeblut, 1981). Эти исследования охватывали широкое множество методик тренировки, проводившейся с учащимися как государственных, так и частных средних школ, и проводились на выборках школьников, принадлежащих к разным слоям населения, включая меньшинства из городских и сельских районов. Общий вывод таков: интенсивное натаскивание в выполнении заданий, сходных с заданиями *SAT*, едва ли приводит к более заметному приросту тестовых показателей по сравнению с тем, который наблюдается в случае повторного проведения *SAT* после года регулярного обучения в средней школе.

Следует также отметить, что Совет колледжей и Совет по проведению письменных экзаменов для аспирантов (*Graduate Record Examination Board*) при создании собственных тестов исследуют новые типы заданий на подверженность тренировке (Evans & Pike, 1973; Powers, 1983; Powers & Swinton, 1984; Swinton & Powers, 1985). Типы заданий, выполнение которых можно заметно улучшить краткосрочными тренировками или узко направленным обучением, исключаются из действующих форм тестов. Ясным примером мог бы служить тип задач, для решения которых требуется простой акт инсайта; стоит испытуемому догадаться, как решить одну такую задачу, он легко справится и со всеми остальными, прямо распространяя на них найденное решение. Если такие задания встретятся испытуемому при последующем тестировании, они будут скорее проверять способность воспроизводить материал по памяти, чем навыки решения задач. Другим примером служат типы сложных заданий, включающих новый или необычный материал и требующих длинных и сложных инструкций (Powers, 1986).

Назначение тренировки (*coaching*) в узком, традиционном смысле этого слова — развить высоко специфичные навыки, которые могут вообще не иметь применения в

реальной жизни, за исключением единичной ситуации тестирования. Подобным же образом, практика «обучения тому, как пройти тест», обычно сосредоточена на конкретной выборке знаний и умений, охватываемых этим тестом, и не затрагивает более широкую область знаний, которую пытаются оценить с помощью данного теста. Так называемые законы о «правдивости в тестировании», или о раскрытии информации, требующие полного доступа к формам теста после его единственного проведения, также способствуют сосредоточению на связанных с конкретным тестом навыках ограниченной применимости. Наконец, поскольку тренировка доступна одним и не доступна другим, она имеет тенденцию вносить индивидуальные различия в строго определенные навыки тестируемых, снижая тем самым диагностическую ценность теста.

Тестовая искушенность. В связи с обсуждаемой проблемой уместно коснуться так называемой тестовой искушенности, или приобретения обширной практики выполнения тестов. При проведении исследований с параллельными формами одного теста обнаруживается тенденция к некоторому повышению результатов второго тестирования. О существенном приросте средних тестовых показателей сообщалось в тех случаях, когда параллельные формы теста предъявляли испытуемым либо непосредственно одна за другой, либо с интервалом, колеблющимся от одного дня до трех лет (Donlon, 1984; Droege, 1966; Peel, 1951, 1952). Сходные результаты были получены на выборках нормальных и интеллектуально одаренных учеников младших классов, учащихся средних школ и колледжей и служащих. Вообще говоря, данные о распределении приростов показателей при повторном тестировании с использованием параллельных форм теста должны приводиться в руководстве к нему, а возможность прироста тестовых показателей в подобных условиях — должны приниматься в расчет при их интерпретации.

Разумеется, круг факторов, вызывающих прирост тестовых показателей, не ограничивается применением параллельных форм теста. Человек, имеющий богатый опыт в выполнении стандартизованных тестов, приобретает тем самым определенные преимущества перед теми, кто впервые участвует в тестировании (Millman, Bishop, & Ebel, 1965; Rodger, 1936). Отчасти эти преимущества вытекают из преодоленного чувства неестественности происходящего, развившейся уверенности в себе и более позитивного отношения к тестовой ситуации. Отчасти же они вызваны некоторым перекрытием содержания и функций большинства тестов. Хорошее знакомство с типами обычных тестовых заданий и практика в заполнении опросных листов также могут несколько улучшить выполнение теста. Особенно важно принимать во внимание тестовую искушенность в случаях, когда сравниваются показатели лиц, опыт которых в прохождении тестирования мог существенно различаться. При компьютерном тестировании следует обращать особое внимание на знакомство тестируемых с этой формой проведения тестов (Hofer & Green, 1985).

Короткие ориентировки и практические занятия могут быть достаточно эффективными при выравнивании тестовой искушенности испытуемых (Anastasi, 1981a; Wahlstrom & Boersman, 1968). Такое ознакомительное обучение по существу ослабляет влияние различий в предшествующем опыте тестируемых. Поскольку эти индивидуальные различия специфичны для конкретной тестовой ситуации, их снижение дало бы возможность более валидной оценки той широкой области поведения, для измерений в которой предназначен определенный тест. Этот подход можно проиллюстрировать на примере издания Совета колледжей «Как пройти SAT I: Тест рассуждений» (*Taking the SAT I: Reasoning Test*) — брошюры, раздаваемой всем абитуриентам, зарегистрировавшимся для прохождения этого теста. Брошюра дает советы по поводу

того, как лучше вести себя во время тестирования, содержит примеры и объяснения различных типов заданий, включенных в тест, и воспроизводит полную форму теста вместе с ключом, рекомендуя учащимся выполнить его за установленное стандартом время и оценить свой результат. Аналогичная брошюра — «Как пройти SAT II: Предметные тесты» (*Taking SAT II: Subject Tests*) — иллюстрирует и объясняет задания из тестов по разным предметам.

Совет по проведению письменных экзаменов для аспирантов (*GRE*) также предоставляет ознакомительные материалы по своему тесту. «Информационный бюллетень» (*Information Bulletin*), распространяемый среди всех поступающих в аспирантуру, дает объяснения образцов заданий из Общего теста (*General Test*) и, кроме того, публикует полную форму теста (с ключом для оценки результатов), проводившегося в прошлом году. Дополнительные формы теста регулярно публикуются в сборнике вариантов GRE: «Подготовка к Общему тесту GRE» (*Practicing to Take the GRE General Test*). Имеются аналогичные практические брошюры, содержащие частные тесты *GRE* по отдельным предметным областям.

Произошедшее в 1980-х и 1990-х гг. увеличение количества ознакомительных материалов по официально проводимым тестам коснулось не только печатной продукции, но и диафильмов, слайдов, микрофильмов, видеокассет и компьютерных программ. Большинство этих материалов было разработано и распространяется Службой тестирования в образовании (*Educational Testing Service*). Некоторые из них предназначены для использования с конкретными тестами, как демонстрационные слайды, входящие в комплект брошюр по *SAT*, и инструкции по интерпретации тестовых показателей *SAT* и тестов достижений Совета колледжей. Компьютерная программа, облегчающая понимание показателей *SAT*, также доступна для всех желающих. Сравнительно сложный пакет обучающих компьютерных программ был разработан для студентов, планирующих пройти Общий тест *GRE*. Благодаря диалоговому режиму работы, этот пакет обеспечивает предъявление пробных заданий, создает условную, нормированную по времени ситуацию тестирования, дает объяснения неправильных ответов на задания и анализ сильных и слабых сторон проходящего тест студента.

Другие материалы (печатная продукция, микрофильмы, мультимедийные комплекты и компьютерные программы) предназначены для более общей ориентации тестируемых, круг которых значительно шире: от учеников младших классов до взрослых. Примером может служить видеодиск «Как самостоятельно подготовиться к стандартизованным тестам» (*On Your Own: Preparing for a Standardized Test*, 1987), созданный для учащихся средних школ, которые могут работать с ним как индивидуально, так и в группах. Другой пример — простое, исчерпывающее руководство в форме книги — «Как пройти тест: Сделай все от себя зависящее!» (*How to Take a Test: Doing Your Best* — Dobbin, 1984). Ряд вспомогательных средств для обеспечения ориентировки лиц, проходящих тестирование, был также подготовлен несколькими крупными коммерческими издательствами тестов и правительственными организациями. Примером последних может служить набор материалов для использования с Батареей тестов общих способностей (*GATB*) Службы занятости США.

Обучение широким когнитивным умениям. Некоторые исследователи пытались найти способы повышения уровня выполнения тестов, продвигаясь в противоположном направлении. Их цель — развитие широко применимых интеллектуальных умений, трудовых навыков и стратегий решения задач. Эффекты такого вмешательства должны, вероятно, проявляться как на уровне тестовых показателей, так и на уровне выбранной в качестве критерия реальной деятельности, например учебной деятельно-

сти в колледже. В соответствии с разграничением, введенным в самом начале этого раздела, предназначение программ этого типа — обеспечить обучение, а не тренировку. В рамках этого направления одни исследователи работали с обучаемыми умственно отсталыми детьми и подростками (Babad & Budoff, 1974; Belmont & Butterfield, 1977; A. L. Brown, 1974; Budoff & Corman, 1974; Campione & Brown, 1979, 1987; Feuerstein, 1979, 1980; Feuerstein, Rand, Jensen, Kaniel, & Tzuriel, 1987). Другие сосредоточились на помощи студентам колледжей и профессиональных школ, имеющим — по разным причинам — существенные пробелы в школьном образовании (Linden & Whimbey, 1990; Whimbey, 1975, 1977, 1980).

Многие из используемых в этих программах методик обучения предназначены для развития эффективной деятельности решения задач (*problem-solving*): обучения (и приучения) тщательно анализировать задачи или вопросы, учитывать все альтернативы, релевантные частности и следствия при поиске решения, взвешенно, а не импульсивно формулировать или выбирать ответы, и применять высокие стандарты при оценивании собственной деятельности. Все это имеет отношение к стратегиям, которые должны улучшить функционирование интеллекта индивидуума не только в ситуации выполнении теста, но также в процессе учебных и многих других повседневных занятий, зависящих от формального обучения. И все же решающим остается вопрос о степени переноса и распространимости таких эффектов на более широкое содержание и разнообразные условия деятельности по сравнению с используемыми в ходе обучения. Результаты, о которых сообщалось до сих пор, выглядят многообещающими. Однако эти программы все еще находятся в стадии опробования, и для установления широты и прочности достигаемых в их рамках эффектов улучшения необходимы дальнейшие исследования.

Краткий обзор. Мы рассмотрели три типа предваряющего тестирование обучения, существенно различающихся по своим целям. Как эти типы обучения влияют на валидность теста и его практическую полезность как оценочного инструмента? Первый тип — тренировка, в смысле интенсивных, многократных упражнений на материале заданий, сходных с заданиями теста. Как отмечалось, в хорошо сконструированных тестах типы заданий отбираются с целью минимизировать чувствительность теста к такому натаскиванию тестируемых; кроме того, в них предусмотрена защита конкретного содержания заданий от несанкционированного доступа. Если даже подобная тренировка и улучшает результаты выполнения теста, соответствующего улучшения в критериальной деятельности обычно не происходит. В связи с этим предваряющая тестирование тренировка может приводить к снижению валидности теста. В результате тест становится менее эффективным средством измерения тех широких способностей, для оценки которых он предназначен, и менее точным средством определения того, обнаружил ли конкретный человек знания и умения, необходимые для успешной деятельности на занимаемом месте.

С другой стороны, ознакомительные мероприятия, ориентирующие испытуемых в основном содержании, процедурах и условиях проведения тестов, имеют целью устранение или выравнивание различий в их опыте прохождения тестов к моменту тестирования. Подобно эффектам тренировки, эти различия представляют собой условия, влияющие на тестовые показатели как таковые, не обязательно сказываясь на более широкой области измеряемого поведения. Следовательно, такие ознакомительные мероприятия должны повышать валидность теста за счет ослабления влияния связанных со спецификой тестирования факторов.

Наконец, практическое обучение широко применимым когнитивным умениям, при условии его эффективности, должно улучшать способность обучаемого справляться с интеллектуальными задачами в последующем. Это улучшение может и должно отражаться на выполнении тестов. Поскольку в результате такого обучения улучшаются как тестовые показатели, так и критериальная деятельность, оно не сказывается на валидности теста, но повышает шансы индивидуума достичь желаемых целей.

Источники информации о тестах

Психологическое тестирование — быстро меняющаяся область. Для нее характерны резкая смена ориентаций, появление новых тестов и обновление старых, непрерывное пополнение данных, которые могут уточнять или полностью изменять интерпретацию оценок по существующим тестам. Ускоряющиеся темпы происходящих в психологическом тестировании перемен, вместе с огромным числом доступных пользователям тестов, делают невозможным обзор конкретных тестов в рамках любого учебника. Более полное и тщательное освещение инструментов тестирования и связанных с ним конкретных проблем можно найти в книгах, посвященных использованию тестов в таких областях, как консультирование, клиническая практика, подбор и расстановка кадров, образование. Ссылки на такие публикации даются в соответствующих главах нашего учебника. Однако, чтобы быть в курсе и не отстать от бурного развития событий в области психологического тестирования, любому, кто работает с тестами, нужно быть знакомым с более прямыми источниками информации о тестах.

Один из самых известных таких источников — «Ежегодник психических измерений» (*Mental Measurements Yearbook*, [*MMY*]), основанный Оскаром К. Буросом и редактируемый им вплоть до 1978 г. С 1985 г. *MMY* стал издавать Институт психических измерений Буроса при Университете штата Небраска. Эта серия ежегодников охватывает почти все доступные для приобретения психологические, образовательные и профориентационные тесты, опубликованные на английском языке. Наиболее полно освещается область бланковых тестов. Каждый выпуск *MMY* включает тесты, опубликованные в течение определенного периода, таким образом дополняя, а не заменяя собой более ранние выпуски. Самые первые публикации в этой серии носили чисто библиографический характер. Начиная с 1938 г., ежегодник приобрел свой нынешний вид и включает критические обзоры тестов, написанные одним или несколькими специалистами, а также полный перечень публикаций по каждому тесту. В дополнение к этому регулярно сообщаются обычные сведения об издателях, цене, формах теста и возрасте лиц, для обследования которых он предназначен. Текущие планы в отношении издания *MMY* — публиковать новый выпуск *MMY* каждые два-три года, издавая к тому же дополнения между двумя очередными выпусками *MMY*.

В наше время статьи о тестах и критические обзоры из *MMY* распространяются в электронном виде через Silver Platter (см. приложение Б). База данных содержит статьи начиная с девятого выпуска *MMY* и обновляется каждые полгода. Еще одно издание Института Буроса — каталог тестов *Tests in Print* — к настоящему времени представлено четвертым томом (*TIP-IV*, 1994) под редакцией L. L. Murphy, Conoley и Impara. Это издание обеспечивает совокупное освещение всех доступных для приобретения англоязычными пользователями тестов, включая фактографическую информацию и перечни ссылок. Каждым последующим изданием *TIP* можно также пользоваться как сводным указателем в отношении всех вышедших ранее выпусков MMY.

Другой важный источник информации об издаваемых тестах — Библиографии собрания тестов (*Test Collection Bibliographies*), подготавливаемые Службой тестирования в образовании (*ETS*). Аннотированные библиографии тестов составляются отдельно по конкретным содержательным областям и обеспечивают исчерпывающее обозрение измерительных инструментов, охватывая все типы тестов, а также тесты, предназначенные для решения специфических задач и обследования особых популяций, таких как лица с физическими недостатками. В каждой статье дается фактографическая информации о конкретном тесте, включая автора, дату издания, издательство, обследуемую совокупность, назначение и все подшкалы теста или измеряемые переменные. Библиографии тестов для отдельных областей можно приобрести за номинальную плату у *Test Collection, ETS* (см. адрес в приложении Б). Это одно из нескольких изданий *ETS*, предоставляющих текущую информацию о тестах и тестировании.

Помимо издаваемых тестов, есть громадное количество некоммерческих тестов, которые описаны или воспроизведены в книгах, журналах или неопубликованных отчетах. Обзоры таких тестов, представляющих интерес главным образом для исследователей, публикуются в различных компендиумах (например, Goldman & Mitchell, 1995). Текущую информацию о некоммерческих тестах можно получить из издания «Тесты на микрофишах» (*Tests in Microfiche*), распространяемого *Test Collection, ETS*. Каждый год база данных *ETS* пополняется новым набором таких тестов, и по запросу можно получить каталог каждого набора. Квалифицированные пользователи имеют возможность приобрести отдельные тесты или их наборы. Краткое и ясное руководство для поиска информации о коммерческих и некоммерческих тестах предоставляется научной дирекцией Американской психологической ассоциации (Finding Information, 1995). Этот источник регулярно обновляется, и по запросу любого желающего автоматически высылается последняя версия руководства.

Что касается пользователей тестов, наиболее прямым источником информации о современных тестах служат каталоги издательств, специализирующихся на выпуске тестов, и руководства по конкретным тестам. Полный перечень таких издательств с указанием их адресов можно найти в последнем выпуске Ежегодника психических измерений. Для облегчения поиска этой справочной информации названия и адреса издательств, чьи тесты упоминаются в нашем учебнике, даны в приложении Б. Каталоги современных тестов можно получить от ведущих издательств по запросу, а квалифицированные пользователи могут приобрести у них комплекты тестов и руководства к ним.

Руководство к тесту должно предоставлять всю информацию, необходимую для проведения теста, подсчета показателей и оценивания его характеристик. Как правило, в нем можно найти полные и подробные инструкции, ключи, нормы и сведения о надежности и валидности. Кроме того, в руководстве к тесту принято указывать количественные и качественные характеристики выборок, на которых устанавливались нормы, надежность и валидность, а также методы вычисления показателей надежности и валидности. В том случае, когда необходимые сведения занимают слишком большой объем и не вписываются в обычно отводимое для них место в руководстве к тесту, в нем должны даваться ссылки на техническое руководство или другие печатные источники, в которых такие сведения можно легко отыскать. Другими словами, руководство должно давать пользователям тестов возможность оценить тест перед тем, как выбрать его для своих конкретных целей. Следует добавить, что некоторые руководства к тестам не оправдывают этих ожиданий. Однако более крупные и ориен-

СТАНДАРТЫ ОБРАЗОВАТЕЛЬНОГО И ПСИХОЛОГИЧЕСКОГО ТЕСТИРОВАНИЯ

Часть I. Технические стандарты конструирования и оценки тестов

1. Валидность.
2. Надежность и ошибки измерения.
3. Усовершенствование и пересмотр тестов.
4. Шкалирование, нормирование, сравнимость и приравнивание показателей.
5. Издание тестов: технические руководства и руководства пользователей.

Часть II. Профессиональные стандарты для пользователей тестов

6. Общие принципы использования тестов.
7. Клиническое тестирование.
8. Образовательное тестирование и психологическое тестирование в школах.
9. Применение тестов в консультировании.
10. Тестирование при приеме на работу.
11. Выдача лицензий и профессиональная аттестация.
12. Оценка программ.

Часть III. Стандарты для специфических контингентов тестируемых

13. Тестирование языковых меньшинств.
14. Тестирование лиц, находящихся в неблагоприятных условиях.

Часть IV. Стандарты проведения тестирования

15. Проведение тестов, получение количественных показателей и их сообщение.
16. Защита прав тестируемых.

Рис. 1-1. Темы, охватываемые Стандартами образовательного и психологического тестирования (AERA, APA, NCME, 1985)

тированные на профессионалов издательства тестов уделяют повышенное внимание подготовке руководств, в полной мере отвечающих научным стандартам. А рост числа подготовленных пользователей тестов служит надежной гарантией того, что такие стандарты будут и дальше поддерживаться и совершенствоваться.

Лаконичные, но исчерпывающие инструкции для оценки психологических тестов можно найти в Стандартах образовательного и психологического тестирования (*Standards for Educational and Psychological Testing*), подготовленных Американской психологической ассоциаций (*APA*) в соавторстве с двумя другими ассоциациями, занимающихся тестированием: Американской ассоциацией педагогических исследований (*American Educational Research Association* [*AERA*]) и Национальным советом по измерениям в образовании (*National Council on Measurement in Education* [*NCME*]). Опубликованные впервые в 1954 г., Стандарты пересматривались в 1966, 1974 и 1985 гг. Следующий всесторонний пересмотр ведется в настоящее время совместными усилиями этих трех ассоциаций.

Потребность ввести Стандарты тестирования[1], которые касались бы не только технического качества тестов, но и влияния тестирования на благополучие человека,

[1] Ради краткости, с этого момента мы будем, следуя общепринятой практике, использовать такое сокращение на протяжении всей книги.

ПРОЕКТ ПЕРЕЧНЯ СТАНДАРТОВ ОБРАЗОВАТЕЛЬНОГО И ПСИХОЛОГИЧЕСКОГО ТЕСТИРОВАНИЯ

Часть I. Конструирование тестов, оценка и документация

1. Валидность.
2. Надежность, ошибки измерения и информационная функция тестовых показателей.
3. Усовершенствование и пересмотр тестов.
4. Шкалирование, нормирование, стандарты и сравнимость показателей.
5. Проведение тестов, получение количественных показателей и их сообщение.
6. Тестовая документация.

Часть II. Честность в тестировании

7. Честность и необъективность.
8. Защита прав тестируемых.
9. Тестирование лиц, для которых английский не является родным языком.
10. Тестирование лиц, неспособных к учебной или трудовой деятельности.

Часть III. Приложения тестирования

11. Общие принципы использования тестов.
12. Психологическое тестирование и оценивание.
13. Образовательное тестирование и оценивание.
14. Тестирование при приеме на работу, выдача лицензий и аттестация.
15. Тестирование в оценке программ и государственной политики.

Рис. 1–2. Темы, выбранные для пересмотренного издания *Стандартов образовательного и психологического тестирования* (AERA, APA, NCME, 1996). Рукопись готовится к изданию. (Воспроизведено с разрешения Объединенного комитета по разработке Стандартов образовательного и психологического тестирования [Dianne Brown, директор проекта])

явно обнаружилась в 1980-е гг. (см. рис. 1–1). То, что эта потребность отражает устойчивую тенденцию, можно заметить по содержанию самого последнего пересмотра Стандартов тестирования. Рис. 1–2 содержит предлагаемый перечень Стандартов, подготовленный комитетом из представителей трех указанных ассоциаций в 1996 г. Очевиден неуклонный рост внимания к тому, чтобы сообразовывать выбор тестов, — а также интерпретацию и использование их показателей, — с доступной информацией об истории жизни тестируемых. Примечательно, что целый раздел Стандартов (часть II) на рис. 1–2 озаглавлен «Честность в тестировании». Пользователи тестов начинают все больше сознавать, что неправильное применение тестов может нанести вред человеку и снизить эффективность его вклада в общество. К тому же широкая и доступная критика неправильного использования тестов, вероятно, в немалой степени содействовала повышению сознательности тех, кто применяет в своей работе тесты, тем самым сокращая число таких случаев. А это, в свою очередь, должно повысить общественное признание потенциальных выгод применения тестов.

2 ИСТОРИЧЕСКИЕ ПРЕДПОСЫЛКИ СОВРЕМЕННОГО ТЕСТИРОВАНИЯ

Краткий обзор исторических предпосылок и истоков психологического тестирования должен создать перспективу и помочь в понимании современных тестов.[1] В свете того, что предшествовало появлению таких тестов, можно яснее увидеть направление развития психологического тестирования в наши дни. Присущие современным тестам отдельные недостатки, равно как и их достоинства, также становятся более понятными при рассмотрении имеющихся в настоящее время измерительных инструментов на фоне исторического прошлого, в котором они берут начало. В этой главе рассматриваются лишь предпосылки и начальный этап развития тестирования как единого целого. Более поздние этапы и линии развития тестирования обсуждаются в последующих главах в связи с конкретными видами тестов, такими как тесты способностей (главы 8–12) или тесты интересов (глава 14), и применением тестов в таких областях, как образование, промышленность, медицина и консультирование (глава 17).

Корни тестирования теряются в древности. Неоднократно сообщалась о системе экзаменов при поступлении на гражданскую службу, существовавшей в китайской империи на протяжении 2000 лет (Bowman, 1989). У древних греков испытание (*testing*) стало неизменным дополнением учебного процесса. Учеников подвергали испытаниям, чтобы оценить, насколько они овладели физическими и умственными навыками (Doyle, 1974). С момента своего появления в средние века европейские университеты при присвоении ученых званий и степеней полагались на результаты официальных экзаменов. Однако, чтобы установить главные события, под влиянием которых сложилось современное тестирование, нет необходимости углубляться в столь от-

[1] Более подробное рассмотрение зарождения тестирования и появления первых психологических тестов можно найти у F. L. Goodenough (1949) и J. Peterson (1926). Что касается общеисторического контекста развития тестирования, см. Boring (1950), G. Murphy and Kovach (1972); более современное изложение истории психологического тестирование дано DuBois (1970) и McReynolds (1975, 1986); у Anastasi (1965) рассмотрены исторические предпосылки изучения индивидуальных различий. Обзор современных тенденций в развитии психологического тестирования также можно найти у Anastasi (1993).

даленное прошлое. Ограничим нашу ретроспективу XIX столетием и рассмотрим важнейшие, с точки зрения развития психологического тестирования, события того времени.

Первые попытки классификации и обучения умственно отсталых

XIX век свидетельствовал о пробуждении устойчивого интереса к гуманному обращению с умственно отсталыми и душевнобольными. До этого времени на долю этих несчастных выпадало пренебрежение, издевательства и даже пытки. Вместе с растущим беспокойством по поводу отсутствия должного ухода за людьми с отклонениями в психике пришло ясное понимание того, что его организация требовала единых критериев для выявления и классификации этих больных. Образование в Европе и США многочисленных общественных заведений по уходу за умственно отсталыми сделало потребность в установлении критериев приема в них и объективной классификации пациентов крайне острой. Прежде всего было необходимо найти способ различать душевнобольных и умственно отсталых. У первых обнаруживались эмоциональные расстройства, не обязательно сопровождавшиеся снижением интеллекта от исходного нормального уровня; вторые характеризовались главным образом интеллектуальным дефектом, врожденным или приобретенным в раннем детстве. По всей вероятности, первая явная формулировка этого дифференциального признака встречается в двухтомном труде французского врача Эскироля (1838), в котором более ста страниц посвящено тому, что теперь принято называть «психическая задержка» (*mental retardation*). Он также указывал на существование множества степеней задержки умственного развития, образующих непрерывный диапазон изменений от нормальности до глубокой идиотии. Пытаясь разработать метод классификации умственной отсталости по форме и степени выраженности, Эскироль опробовал несколько способов и пришел к выводу, что способность индивидуума пользоваться языком есть самый надежный критерий его интеллектуального уровня. Примечательно, что используемые в наше время критерии задержки умственного развития также являются преимущественно лингвистическими, а современные тесты интеллекта сильно насыщены вербальным содержанием. Та важная роль, которую вербальная способность играет в нашем понятии интеллекта, будет неоднократно продемонстрирована в последующих главах.

Особое значение имеет вклад другого французского врача — Сегена, первым начавшего обучать умственно отсталых. Отвергнув преобладавшее в то время мнение о неизлечимости умственной отсталости, Сеген (1866–1907) много лет опробовал метод обучения, названный им физиологическим, и в 1837 г. основал первую школу для обучения умственно отсталых детей. В 1848 г. он эмигрировал в Америку, где его идеи получили широкое признание. Многие из методик тренировки органов чувств и мышечного аппарата, используемых в настоящее время в учреждениях для умственно отсталых, были изобретены Сегеном. Эта методики позволяют проводить с глубоко отсталыми детьми интенсивные занятия по сенсорному различению и развитию моторного контроля. Некоторые из приемов, разработанных с этой целью Сегеном, были со временем включены в состав практических или невербальных тестов интеллекта.

Как пример можно назвать Доску форм Сегена (*Seguin Form Board*), при использовании которой в качестве диагностического инструмента от индивидуума требуется как можно быстрее вставить фигуры разной формы в соответствующие им углубления.

Более чем полвека спустя после работ Эскироля и Сегена французский психолог Альфред Бине убеждал чиновников и общественность в том, чтобы детей, не справляющихся с обучением в обычной школе, прежде чем отчислять, обследовали и, если они будут признаны обучаемыми, направляли в специальные классы (Т. Н. Wolf, 1973). Вместе с другими членами Общества психологического изучения ребенка (*Society for the Psychological Study of the Child*) Бине побуждал Министерство общественного образования (*Ministry of Public Instruction*) предпринять шаги к улучшению положения умственно отсталых детей. Конкретным результатом стало создание министерской комиссии по изучению отсталых детей, в состав которой был включен и Бине. Какую роль это назначение сыграло в истории психологического тестирования, расскажем несколько позднее.

Первые психологи-экспериментаторы

Стоящих у истоков экспериментальной психологии ученых XIX в. вообще не интересовало измерение индивидуальных различий. Главной целью психологов того периода было составление обобщенных описаний человеческого поведения. Поэтому их внимание было приковано не к различиям в поведении, а к его единообразию. Индивидуальные различия либо игнорировали, либо воспринимали как неизбежное зло, ограничивающее применимость обобщений. Таким образом, сам факт, что два человека, наблюдаемые в идентичных условиях, реагировали на эти условия по-разному, рассматривался этими психологами как разновидность погрешности. Наличие такой погрешности, или индивидуальной изменчивости, превращало обобщения из точных в приближенные. Подобное отношение к индивидуальным различиям господствовало в таких научных лабораториях, как лаборатория Вундта, основанная им в 1879 г. в Лейпциге, где обучались многие из первых психологов-экспериментаторов.

На выборе тем, как и на многих других сторонах работы основателей экспериментальной психологии, сказывалось влияние их профессиональной подготовки в области физиологии и физики. Проблемы, исследовавшиеся ими в лабораториях, в основном касались чувствительности к зрительным, слуховым и другим сенсорным стимулам и времени простой реакции. Как станет видно из последующих разделов, этот акцент на сенсорных феноменах нашел отражение и в характере первых психологических тестов.

Экспериментальная психология XIX в. повлияла на направление развития тестирования еще в одном отношении. Первые психологические эксперименты выявили необходимость строгого контроля условий проведения наблюдений. Например, формулировка инструкций, дававшихся испытуемому в эксперименте на время реакции, могла существенно увеличить или уменьшить ее скорость. Опять-таки, яркость или цвет окружающего фона могли заметно повлиять на восприятие зрительного стимула. Тем самым была ясно доказана важность проведения наблюдений за реакциями всех испытуемых в стандартизованных условиях. Со временем такая стандартизация процедуры проведения исследования стала одним из отличительных признаков психологических тестов.

Вклад Френсиса Гальтона

Именно благодаря научной деятельности английского биолога Френсиса Гальтона развитие тестирования как самостоятельного направления стало набирать темпы. Его многочисленные и разнообразные исследования объединял интерес к наследственности человека. В процессе этих исследований Гальтон пришел к пониманию необходимости количественного измерения характеристик людей, состоящих и не состоящих в родстве. Только таким путем он мог установить, например, точную степень сходства между родителями и потомками, братьями и сестрами, родными и двоюродными, или близнецами. Преследуя эту цель, Гальтон способствовал созданию ряда образовательных учреждений, в которых вел систематические антропометрические измерения учащихся. Он также организовал на Всемирной выставке 1884 г. антропометрическую лабораторию, где за три пенса посетители могли измерить некоторые из своих физических характеристик и пройти тесты на остроту зрения и слуха, мышечную силу, время реакции и другие элементарные сенсомоторные функции. После закрытия выставки лаборатория была переведена в Южно-Кенсингтонский музей в Лондоне и действовала там еще шесть лет. Такими методами постепенно накапливались первые систематические данные об индивидуальных различиях в простых психологических процессах.

Гальтон сам разработал большинство простых тестов, применявшихся в его антропометрической лаборатории, и многие из них еще знакомы нам либо в своем оригинальном, либо в модифицированном виде. В качестве примеров можно назвать линейку Гальтона для зрительного различения длины, свисток Гальтона для определения верхнего частотного порога слуховых ощущений и градуированную серию разновесов для измерения кинестетического различения. Гальтон полагал, что тесты сенсорного различения могут служить средством измерения интеллекта человека. В этом отношении на него отчасти повлияло учение Джона Локка. Так, Гальтон писал: «Информация о внешних событиях поступает к нам только от наших органов чувств, и чем лучше эти органы улавливают различия, тем обширнее поле, на котором могут действовать наши интеллект и рассудок» (Galton, 1883, p. 27). Гальтон также отметил, что при крайней степени слабоумия нарушается способность различать тепло, холод и боль. Это наблюдение только усилило его убеждение в том, что различительная способность органов чувств «в целом должна быть самой высокой у наиболее интеллектуально одаренных» (Galton, 1883, p. 29).

Френсис Гальтон также был пионером в применении оценочных шкал, методов анкетирования и методики свободных ассоциаций, впоследствии использовавшихся для самых различных целей. Еще одной заслугой Гальтона по праву считают разработку методов математической статистики для анализа данных об индивидуальных различиях. Он отобрал и упростил ряд вычислительных процедур, выведенных математиками. Гальтон придал этим процедурам такую форму, чтобы ими мог воспользоваться исследователь, не имеющий математической подготовки, при желании количественно обработать результаты тестов. В этом направлении продолжали работать многие из учеников Гальтона, среди которых наиболее выдающимся был Карл Пирсон.[1]

---

[1] Увлекательное изложение истории развития основных статистических понятий и биографий причастных к этому ученых см. у Cowles (1989).

Джеймс Кэттелл и первые «умственные тесты»

Особо заметное место в развитии психологического тестирования занимает американский психолог Джеймс Маккин Кэттелл, работы которого объединили недавно возникшую экспериментальную психологию с еще более молодым направлением — тестированием. В Лейпциге, вопреки неприятию В. Вундтом такого типа исследований, Кэттелл написал диссертацию об индивидуальных различиях во времени реакции. Позднее, когда в 1888 г. он читал лекции в Кембридже, его интерес к измерению индивидуальных различий усилился благодаря влиянию Гальтона. По возвращении в Америку Кэттелл активно занялся созданием лабораторий экспериментальной психологии и распространением тестирования.

Термин «умственный тест» (*mental test*) впервые появился в психологической литературе в статье Кэттелла, опубликованной в 1890 г. В этой статье описана серия тестов, ежегодно проводившихся для определения интеллектуального уровня студентов колледжей. Эти тесты должны были проводиться индивидуально и включали измерения мышечной силы, скорости движения, чувствительности к боли, остроты зрения и слуха, различения веса, времени реакции, памяти и т. п. При выборе своих тестов Кэттелл был солидарен с Гальтоном в том, что оценку интеллектуальных функций можно получить посредством тестов сенсорного различения и времени реакции. Предпочтение таких тестов Кэттеллом объяснялось и тем фактом, что простые функции могли быть измерены с большой точностью, а разработка объективных методов измерения более сложных функций казалась в то время совершенно безнадежной задачей.

Задания, подобные тестам Кэттелла, можно было обнаружить практически в любой из многих серий тестов, разработанных в последнее десятилетие XIX в. Такие серии тестов проводили на школьниках, студентах колледжей и смешанных выборках взрослых. На Колумбийской выставке, проходившей в 1893 г. в Чикаго, Ястров выставил стенд, к которому приглашал посетителей проверить свои сенсорные, моторные и простые перцептивные процессы и сравнить свои достижения с нормами (J. Peterson, 1926; Philippe, 1894). Немногочисленные попытки оценить эти первые тесты дали обескураживающие результаты. Сопоставление результатов по двум тестам у одного и того же человека практически не обнаружило сколько-нибудь существенного соответствия между ними (Sharp, 1898–1899; Wissler, 1901); не удалось также выявить никакой связи результатов тестирования с независимыми оценками интеллектуального уровня, основанными на суждениях учителей (T. L. Bolton, 1891–1892; J. A. Gilbert, 1894) или с академической успеваемостью (Wissler, 1901).

Некоторые серии тестов, составленные в это время европейскими психологами, предусматривали также измерение более сложных функций. Немецкий психолог Э. Крепелин (1895), которого прежде всего интересовало клиническое обследование пациентов с психическими расстройствами, создал большую серию тестов для измерения того, что он считал основными факторами при описании характера индивидуума. Эти тесты, в основном использовавшие элементарные арифметические операции, предназначались для измерения эффектов упражнения, памяти, подверженности утомлению и отвлечению внимания. Другой немецкий психолог, Г. Эббингауз (1897), проводил со школьниками тесты на арифметический счет, сохранение заученного материала в памяти и завершение предложений. Наиболее сложный из этих трех тестов —

тест на завершение предложений — оказался единственным, обнаружившим явное соответствие учебным достижениям детей.

В статье, опубликованной во Франции в 1895 г., А. Бине и В. Анри раскритиковали большинство имевшихся в наличии серий тестов за неоправданно большое внимание к сенсорным характеристикам и элементарным специальным способностям. Кроме того, они утверждали, что при измерении более сложных функций большой точности не требуется, поскольку в этих функциях индивидуальные различия особенно велики, и предлагали обширный перечень разнообразных тестов, предназначенных для измерения таких функций, как память, воображение, внимание, понимание, внушаемость, эстетическое восприятие, и многих других. Уже в этих тестах можно заметить те тенденции, которые в конце концов привели к созданию известных шкал интеллекта Бине.

А. Бине и появление тестов интеллекта

Бине и его сотрудники много лет посвятили активным и оригинальным исследованиям способов измерения интеллекта. Были испробованы многие подходы, включая даже измерение формы черепа, лица, рук и анализ почерка. Результаты, однако, все более убеждали, что непосредственное, хотя бы и грубое, измерение сложных интеллектуальных функций наиболее перспективно. И наконец, одна неожиданная ситуация привела Бине к желанной цели. В 1904 г. министр общественного образования назначил Бине в уже упоминавшуюся Комиссию по изучению методов обучения умственно отсталых детей. Именно в связи с целями, стоящими перед этой комиссией, Бине в сотрудничестве с Симоном создал первую шкалу Бине—Симона (Binet, & Simon, 1905).

Эта шкала, известная нам как шкала 1905 г., состояла из 30 заданий, или тестов, расположенных по возрастающей трудности. Уровень трудности определялся эмпирически, путем проведения этих тестов на 50 нормальных детях в возрасте от 3 до 11 лет, а также на нескольких умственно отсталых детях и взрослых. Тесты предназначались для измерения широкого круга функций, с особым акцентом на способностях к суждению, пониманию и рассуждению, которые Бине считал основными компонентами интеллекта. Хотя сенсорные и перцептивные тесты также входили в эту шкалу, в ней, по сравнению с большинством серий тестов того времени, существенно возросла доля вербального материала. Шкалу 1905 г. ее создатели представили как предварительный, пробный образец измерительного инструмента, и пока им не удалось найти строгий объективный метод получения совокупного, общего показателя из множества результатов по отдельным тестам.

Во втором варианте шкалы, редакция 1908 г., общее число тестов было увеличено, некоторые неудачные тесты более ранней шкалы изъяты, и все тесты были сгруппированы по возрастным уровням на основе их выполнения примерно 300 нормальными детьми в возрасте от 3 до 13 лет. Так, к уровню 3 лет были отнесены все тесты, с которыми справлялось от 80 до 90 % нормальных трехлетних детей; к уровню 4 лет — все тесты, с которыми справлялось столько же нормальных четырехлетних детей, и т. д. до 13 лет. Показатель ребенка по всем тестам можно было в этом случае выразить в виде *умственного уровня*, соответствующего возрасту нормальных детей, результатов которых он достигал. В разных переводах и переработках шкал Бине термин «ум-

ственный уровень» обычно заменялся термином «умственный возраст», поскольку умственный возраст — понятие простое и доступное, и его введение несомненно способствовало популяризации интеллектуального тестирования.[1] Сам А. Бине, однако, избегал термина «умственный возраст» из-за вытекающих из него, но, увы, необоснованных следствий о нормах возрастного развития и предпочитал более нейтральный термин «умственный уровень» (T. H. Wolf, 1973).

Третий вариант шкалы Бине—Симона появился в 1911 г., отмеченном преждевременной смертью Альфреда Бине. Шкала эта по сравнению с предыдущей претерпела незначительные изменения, которые свелись к перестановке отдельных тестов, добавлению новых тестов для некоторых возрастных уровней и расширению верхней границы шкалы до уровня взрослого человека.

Еще до пересмотра 1908 г. тесты Бине—Симона привлекли широкое внимание психологов всего мира. Их переводы и адаптации появились во многих странах, включая США, где было опубликовано несколько вариантов этой шкалы. Первый вариант был подготовлен Г. Г. Годдардом (H. H. Goddard), работавшим в то время психологом-исследователем в Вайнлендской исправительной школе (для умственно отсталых детей). Шкала Бине—Симона в редакции Годдарда оказала решающее влияние на принятие тестирования интеллекта медицинскими работниками (Zenderland, 1987). Она появилась в благоприятный момент, удовлетворив настоятельную потребность специалистов в стандартизованной мерке для постановки диагноза и классификации лиц с задержкой умственного развития. Однако в качестве инструмента тестирования эта шкала вскоре была вытеснена более широкой и совершенной в психометрическом отношении шкалой умственного развития Стэнфорд—Бине, разработанный под руководством Л. М. Тёрмена в Стэнфордском университете (Terman, 1916). Именно в этом варианте шкалы был впервые использован коэффициент интеллекта (*IQ*), или отношение умственного возраста к хронологическому. Последующие редакции этой шкалы получили широкое применение и будут более основательно рассмотрены в главе 8. Особый интерес представляет также первая редакция шкалы Бине—Симона, произведенная Ф. Кюльманом, в которой нижняя возрастная граница была снижена до 3 мес. (Kuhlmann, 1912). Шкала Кюльмана—Бине представляет собой одну из самых ранних попыток разработать тесты интеллекта для младенцев и дошкольников.

Групповое тестирование

Тесты Бине, как и все их редакции, являются *индивидуальными шкалами* в том смысле, что они могут проводиться только с одним человеком за раз. Большинство тестов в этих шкалах требуют от испытуемого устного ответа или манипулирования

---

[1] Ф. Л. Гудинаф (F. L. Goodenough, 1949, p. 50–51) отмечает, что в 1887 г., за 21 год до появления шкалы Бине—Симона 1908 г., С. Э. Шайе (S. E. Chaille) опубликовал в Нью-Орлеанском медицинском журнале (New Orleans Medical and Surgical Journal) серии тестов для младенцев, распределив их в соответствии с возрастом, в котором малыши обычно справляются с этими тестами. Частично вследствие малой доступности журнала, частично же из-за того, что ученый мир еще не был к этому готов, идея возрастной шкалы в то время осталась незамеченной. На создание подобной шкалы самим А. Бине повлияли работы некоторых его современников, особенно Блин (Blin) и Даме (Damaye), составивших серию устных вопросов, из ответов на которые они выводили единый суммарный показатель для каждого ребенка (T. H. Wolf, 1973).

стимульным материалом, причем в некоторых из них нужно учитывать индивидуальное время выполнения задания. По этим и другим причинам такие тесты не приспособлены для группового использования. Для тестов типа шкалы Бине характерно и то, что проводить их может только квалифицированный специалист. Такие тесты по существу являются клиническими инструментами, приспособленными для интенсивного изучения индивидуальных случаев.

Групповое тестирование, так же как и первая шкала Бине, было создано в ответ на настоятельную потребность практики. Когда Соединенные Штаты вступили в Первую мировую войну в 1917 г., Американская психологическая ассоциация учредила комитет для рассмотрения тех средств, которыми психология могла бы помочь ведению войны. Этот комитет под руководством Роберта М. Йеркса выявил потребность в быстрой классификации полутора миллионов новобранцев по их уровню общего интеллекта. Такая информация имела значение для многих административных решений, включая признание негодными к военной службе, распределение по родам войск, прием в лагеря обучения офицеров и т. п. Для решения этой задачи военные психологи привлекли все имеющиеся тестовые материалы, в частности неопубликованный тест для группового тестирования интеллекта, подготовленный Артуром С. Отисом и специально переделанный им для потребностей армии. Основным достоинством теста Отиса, который он составил еще во время обучения в аспирантуре у Л. М. Тёрмена, было введение задач с множественным выбором ответов и других типов «объективных» заданий.

Тесты, которые в конце концов создали военные психологи, стали называться *армейский альфа* (*Army Alpha*) и *армейский бета* (*Army Beta*). Первый предназначался для общего обычного тестирования; второй представлял собой невербальную шкалу, рассчитанную на неграмотных и новобранцев иностранного происхождения, которые не могли пройти тестирование на английском языке. Оба теста пригодны для проведения в больших группах людей.

Вскоре после окончания Первой мировой войны было получено разрешение использовать военные тесты в гражданских целях. Армейские альфа и бета тесты не только сами неоднократно перерабатывались, но и послужили образцом для многих групповых тестов интеллекта. Тестирование как самостоятельное направление сделало гигантский скачок в своем развитии. Вскоре были разработаны групповые тесты интеллекта для лиц всех возрастов и категорий — от дошкольников до аспирантов. Еще совсем недавно невыполнимые, массовые программы тестирования затевались с завидным оптимизмом. Поскольку групповые тесты создавались как средства массового тестирования, их инструкции и процедура проведения были достаточно просты и потому предъявляли минимум требований к подготовке лиц, работающих с такими тестами. Школьные учителя начали проводить тесты интеллекта в своих классах. Студенты колледжей перед зачислением проходили стандартную проверку. Предпринималось широкое обследование особых групп взрослого населения, таких как заключенные. И скоро широкая публика превратилась в «*IQ*-сознающую».

Применение таких групповых тестов интеллекта значительно обогнало ход их технического усовершенствования. В стремлении собрать как можно больше «объективных» данных о людях и извлечь из этих данных практические выгоды часто забывалось, что тесты все еще были технически несовершенными инструментами. Когда же подобные тесты не оправдывали необоснованных ожиданий, это приводило к скепти-

цизму и неприязни в отношении тестирования вообще. Таким образом, тестовый бум 1920-х гг., основанный на неразборчивом использовании тестов, по-видимому, столько же мешал, сколько и способствовал прогрессу психологического тестирования.

Тестирование способностей

Хотя тесты интеллекта изначально задумывались как инструменты, позволяющие брать пробы широкого множества функций для того, чтобы оценить общий интеллектуальный уровень индивидуума, вскоре стало очевидным, что они обладают весьма ограниченной зоной охвата, в которую не попал ряд важных функций. Фактически, большинство тестов интеллекта в основном измеряло вербальную способность и, в несколько меньшей степени, способность оперировать числовыми и другими абстрактными и символическими отношениями. Постепенно психологи пришли к признанию того, что термин «тест интеллекта» искажает истинное положение вещей, поскольку такие тесты измеряли только некоторые аспекты интеллекта.

Несомненно, эти тесты охватывали способности, имеющие первостепенное значение в той культуре, для которой их разрабатывали. Но стало ясно, что было бы предпочтительнее подыскать для них более точные названия, исходя из типа той информации, которую они могут давать. Например, ряд тестов, называвшихся в 1920-х гг. тестами интеллекта, позднее стали называть тестами академических способностей. Такая смена терминологии была вызвана осознанием того, что многие так называемые тесты интеллекта на самом деле измеряют комбинацию способностей, востребуемых и развиваемых учебной деятельностью.

Еще до Первой мировой войны психологи пришли к пониманию необходимости дополнить тесты общего интеллекта тестами специальных способностей. *Тесты специальных способностей* разрабатывались преимущественно для использования в профориентации, а также при отборе и распределении промышленного и военного персонала. Самыми распространенными среди них были тесты технических, конторских, музыкальных и художественных способностей.

Критическая оценка тестов интеллекта, последовавшая за их необычно широким распространением и неразборчивым использованием, выявила еще один заслуживающий внимания факт: выполнение конкретным человеком разных частей такого теста обнаружило заметную вариацию. Это особенно ясно проявилось в групповых тестах, в которых задания обычно подразделяются на субтесты относительно однородного содержания. Так, человек мог иметь относительно высокий показатель по вербальному и низкий по числовому субтесту, или наоборот. В какой-то степени такая внутренняя вариабельность наблюдается и в тестах типа Стэнфорд—Бине, в которых для конкретного человека могут оказаться трудными, например, задания, содержащие слова, а выигрышными задания, использующие картинки или геометрические фигуры.

Пользователи тестов, и особенно клиницисты, часто прибегали к сравнению выполнения обследуемым разных частей теста для того, чтобы глубже проникнуть в его психологическую конституцию. Таким образом не только *IQ* или какой-то другой общий показатель, но и результаты выполнения группы заданий или субтестов учитывались при анализе индивидуальных случаев. Однако такая практика пригодна не всегда, поскольку тесты интеллекта не рассчитаны на дифференциальный анализ способностей. Часто сравниваемые субтесты содержат слишком мало заданий, чтобы дать

устойчивую или надежную оценку той или иной специальной способности. В результате, различия между показателями по отдельным субтестам у конкретного человека нередко изменяются на противоположные при его повторном обследовании в другой день с помощью того же теста (или параллельной формы такого теста). Чтобы осуществлять такие внутрииндивидные сравнения, необходимы тесты, специально предназначенные для выявления различий в работе анализируемых функций.

В то время как практическое применение тестов способствовало осознанию потребности в комплексных тестах способностей, одновременное развитие исследований структуры черт индивидуума постепенно снабжало ученых средствами для конструирования таких тестов. В статистических исследованиях природы интеллекта выявлялись взаимосвязи показателей по широкому кругу тестов, которые проводились на больших выборках испытуемых. Такие исследования были начаты английским психологом Чарльзом Спирменом (1904, 1927) в первом десятилетии XX в. В результате последующей разработки и усовершенствования методов этого направления в трудах английских и таких американских психологов, как Т. Л. Келли (T. L. Kelly, 1928) и Л. Л. Терстоун (L. L. Thurstone, 1938, 1947b), сложилась группа методов, получивших название *факторного анализа.*

Вклад методов факторного анализа в конструирование тестов будет более полно рассмотрен в главе 11. Сейчас достаточно отметить, что данные, полученные с его помощью, показали наличие ряда относительно независимых факторов, или черт. Некоторые из этих черт были в той или иной мере представлены в традиционных тестах интеллекта. Примерами такого вида черт могут служить вербальное понимание и числовое рассуждение. Черты другого вида, такие как пространственные, перцептивные и механические способности, чаще выявлялись не тестами интеллекта, а тестами специальных способностей.

Один из главных практических результатов применения факторного анализа — разработка *комплексных батарей способностей,* предназначенных для измерения степени выраженности у индивида каждой из входящих в установленный набор черт. Вместо общего показателя, или *IQ,* в этом случае получают отдельные оценки таких черт, как вербальное понимание, способность к счету в уме, пространственное воображение, арифметическое рассуждение и скорость восприятия. Такие батареи оказались подходящим инструментом для внутрииндивидного анализа, или дифференциальной диагностики, — желанной цели, которую пользователи тестов в течении многих лет пытались реализовать на основе приблизительных и часто ошибочных результатов тестов интеллекта. Кроме того, эти батареи в составе полной программы тестирования дают значительный объем информации, получаемой ранее только с помощью тестов специальных способностей, поскольку в зону охвата комплексных батарей способностей попадают некоторые черты, обычно не оцениваемые тестами интеллекта.

Комплексные батареи способностей представляют собой относительно позднее достижение в области тестирования. Почти все они появились после 1945 г. В этой связи следует отметить труд военных психологов во время Второй мировой войны. Большинство тестовых исследований, проводившихся в вооруженных силах, основывалось на применении факторного анализа и было нацелено на создание комплексных батарей способностей. В военно-воздушных силах, например, специальные батареи конструировались для пилотов, бомбардиров, радистов, штурманов и многих других военных специалистов. Отчет об одних только тестовых батареях, подготовленных в ВВС, занимает по меньшей мере 9 из 19 томов, посвященных программе авиационной

психологии во время Второй мировой войны (*Army Air Forces*, 1947—1948). Аналогичным образом был разработан ряд комплексных батарей способностей для использования в гражданской сфере, и они широко применяются в образовательном и профессиональном консультировании, а также при отборе и распределении персонала. Примеры таких батарей будут рассмотрены в главе 10 и 17.

Более современная тенденция развития, обнаружившаяся в конце 1980-х — начале 1990-х гг., обеспечивает принципиальную интеграцию двух ранее противостоящих подходов к психическому измерению, представленным традиционными тестами интеллекта и комплексными батареями способностей (Anastasi, 1994). Наступает понимание того, что способность человека можно адекватно оценивать на разных уровнях широты, от узко определяемых специальными тестами (или даже отдельными заданиями) способностей через все более широкие уровни черт до полной оценки, такой как традиционный *IQ*. Различным целям тестирования лучше всего соответствуют разные уровни широты. Поэтому недавно разработанные тесты интеллекта, такие как Дифференциальные шкалы способностей (*Differential Ability Scales*), или современные версии более ранних тестов, такие как четвертая редакция шкалы Стэнфорд—Бине, сочетают широкий охват разнообразных способностей с гибкой многоуровневой системой подсчета показателей соответственно конкретным целям тестирования. Хотя оба этих примера относятся к индивидуальным тестам интеллекта, тот же комплексный и гибкий подход к конструированию и проведению тестов реализуется при создании групповых тестовых батарей, таких как рассматриваемые в главе 10. Теоретическая основа и практические следствия такого слияния программ тестирования способностей обсуждаются в главе 11, в связи с современными достижениями в области изучения природы интеллекта.

Стандартизованные тесты достижений

Между тем как психологи занимались разработкой тестов интеллекта и способностей, традиционные школьные экзамены подвергались некоторым техническим усовершенствованиям (O. W. Caldwell, & Courtis, 1923; Ebel, & Damrin, 1960). Важный шаг в этом направлении был сделан общественными школами Бостона, заменившими в 1845 г. устные опросы учащихся специально приглашаемыми экзаменаторами на письменные экзамены. Главные аргументы, выдвигавшиеся тогда в защиту этого нововведения, сводились к тому, что письменные экзамены ставят всех учеников в равное положение, позволяют охватить большее содержание, уменьшают элемент случайности в выборе задаваемого вопроса и сводят на нет возможную необъективность экзаменатора. Все эти аргументы звучат на удивление знакомо, так как значительно позднее они использовались для обоснования замены в тестах вопросов, предполагающих свободные, описательные ответы, на объективные задания с множественным выбором.

С наступлением XX столетия начали появляться первые стандартизованные тесты для измерения результатов школьного обучения. Под влиянием пионерских работ Э. Л. Торндайка (E. L. Thorndike) в этих тестах использовались принципы измерения, разработанные в психологических лабораториях. В качестве примера таких тестов можно назвать шкалы оценки качества почерка и письменных сочинений, а также тесты на правописание и решение арифметических примеров и задач. Несколько позднее стали появляться батареи достижений, начало которым было положено изданием

в 1923 г. первой редакции Стэнфордского теста достижений (*Stanford Achievement Test*). Его авторами были три ведущих специалиста того времени в области разработки тестов: Трумэн Л. Келли (Truman L. Kelley), Джайлс М. Рач (Giles M. Ruch) и Льюис М. Тёрмен (Lewis M. Termen). Отвечая многим требованиям современного тестирования, эти батареи обеспечивали сопоставимые меры выполнения заданий по разным школьным предметам, оцениваемого относительно одной нормативной группы.

К этому времени стали очевидными разногласия среди учителей в оценке результатов описательных тестов. К 1930 г. было признано, что описательные тесты по сравнению с объективными заданиями[1] «нового типа» не только отнимают у экзаменаторов и экзаменующихся больше времени, но и дают менее надежные результаты. Чем шире применялись объективные задания в стандартизованных тестах достижений, тем больше значения придавалось им при разработке заданий для тестов на понимание и применение знаний и других широких образовательных целей. Четвертое десятилетие XX в. отмечено также внедрением машин для подсчета тестовых показателей, и новые объективные тесты прекрасно подходили для автоматизированной обработки.

Создание местных, региональных и национальных программ тестирования было еще одной параллельной линией развития, заслуживающей упоминания. Вероятно, наибольшей известностью пользуется программа Совета по вступительным экзаменам в колледжи (*College Entrance Examination Board* [*CEEB*]). Принятая в начале XX в. с целью уменьшить дублирование экзаменов для поступающих в колледжи, эта программа претерпела глубокие изменения в том, что касается процедур тестирования, а также числа и типов участвующих в ней колледжей. Эти изменения отражали события переходного периода, связанные с развитием тестирования и становлением системы образования. В 1947 г. функции проведения тестирования, распределенные между Советом по вступительным экзаменам в колледжи (*CEEB*), корпорацией Карнеги (*Carnegie Corporation*) и Американским управлением образования (*American Council on Education*), были переданы вновь созданной Службе тестирования в образовании (*ETS*), со временем принявшей на себя ответственность за все программы тестирования для университетов, профессиональных училищ, правительственных учреждений и других организаций. Следует также упомянуть программу тестирования американских колледжей (*American College Testing Program* [*ACT Program*]), созданную в 1959 г. для отбора поступающих в колледжи, не охваченные программой *CEEB*, и несколько национальных программ тестирования для отбора высокоодаренных учащихся с целью присуждения поощрительных стипендий.

Тесты достижений используются не только в сфере образования, но и при отборе поступающих на работу в промышленность и государственные учреждения. Как уже отмечалось, систематические экзамены при приеме на гражданскую службу в китайской империи были введены примерно за 150 лет до наступления нашей эры. В европейских странах отбор правительственных служащих на основе экзаменов был введен в конце XVIII – начале XIX в. Комиссия гражданской службы США утвердила обязательные конкурсные экзамены в 1883 г. (Kavruck, 1956). Методы составления тестов, разработанные до и во время Первой мировой войны, были внедрены в экзаменационную программу государственной гражданской службы США после назначения Л. Дж. О'Рурке (L. J. O'Rourke) директором созданного в 1922 г. исследовательского

[1] Исследования, касающиеся относительной эффективности «свободных» и «объективных» типов заданий, приведены в главе 17 в связи с использованием тестов в сфере образования.

отдела. В наши дни эту работу проводит большая и хорошо технически оснащенная научно-исследовательская группа в составе Службы управления кадрами США (*U. S. Office of Personnel Management*).

По мере того как все больше психологов, имеющих психометрическую подготовку, участвовали в создании стандартизованных тестов достижений, технические аспекты этих тестов приобретали все большее сходство с техническими аспектами тестов интеллекта и способностей. Методики конструирования и оценивания всех этих тестов имеют много общего. Усиливающееся стремление создать тесты достижений, которые бы действительно измеряли достижение человеком основных целей образования, а не просто оценивали объем заученных конкретных сведений, способствовало тому, что и содержание тестов достижений становилось все больше похожим на содержание тестов интеллекта. В настоящее время разница между этими двумя типами тестов, в основном, сводится к различиям в степени специфичности содержания и необходимости изучения определенной дисциплины до прохождения тестирования.

Оценка личности

Еще одна область психологического тестирования, которую мы будем обсуждать в главе 13–16, имеет дело с аффективными, или неинтеллектуальными, аспектами поведения. Предназначенные для этого тесты обычно называют тестами личности, хотя некоторые психологи используют термин «личность» более широко, для указания на целостного человека. В последнем случае оценка личности включала бы как интеллектуальные, так и неинтеллектуальные черты человека. Однако в психологическом тестировании термин «тест личности» чаще всего относится к средствам измерения таких индивидуальных особенностей, как эмоциональные состояния, межличностные отношения, мотивация, интересы и аттитюды.

Примером первых попыток тестирования личности может служить использование Крепелином теста свободных ассоциаций в работе с душевнобольными. В этом тесте обследуемому человеку предъявляются специально подобранные слова-стимулы, на которые он должен отвечать первым пришедшим в голову словом. Крепелин (Kraepelin, 1892) использовал эту же методику для изучения психологических эффектов утомления, голода и приема лекарственных препаратов и пришел к выводу, что все эти факторы увеличивают относительную частоту поверхностных ассоциаций. Примерно в эти же годы Р. Соммер (Sommer, 1894) высказал предположение, что тест свободных ассоциаций можно было бы использовать для дифференциальной диагностики психических расстройств. Впоследствии технику свободных ассоциаций стали использовать для самых разных целей тестирования, она не теряет своего значения и в наши дни. Здесь следует упомянуть вклад Ф. Гальтона, К. Пирсона и Дж. Кэттелла в разработку стандартизованных опросников и рейтинговых шкал. Хотя первоначально эти методики разрабатывались с совершенно иными целями, со временем они стали использоваться другими исследователями при конструировании ряда наиболее распространенных типов современных тестов личности.

Прототипом современных личностных опросников, или *вопросников самоотчета* (*self-report inventory*), обычно считают Бланк личных сведений (*Personal Data Sheet*), разработанный Р. Вудвортсом в годы Первой мировой войны (DuBois, 1970; Franz, 1919, p. 171–176; L. R. Goldberg, 1971; Symonds, 1931, chap. 5). Этот тест задумывался

как грубый метод выявления и отсеивания с военной службы лиц с серьезными психическими нарушениями. Он состоял из набора вопросов, касающихся типичных психопатологических симптомов, на которые отвечали сами респонденты. Общий показатель получался путем подсчета отмеченных у себя симптомов. Во время войны Бланк личных сведений так и не был доведен до уровня практического использования, но сразу же после ее окончания были подготовлены его формы для гражданского использования, в том числе специальная форма для опроса детей. Кроме того, Бланк личных сведений Вудвортса послужил образцом для последующей разработки большинства инвентарей эмоциональных приспособительных реакций. В некоторых из них делались попытки подразделить эти эмоциональные реакции на ряд специфических форм в зависимости от приспособления к домашней, учебной или рабочей обстановке. В других же упор делался на более узкой сфере поведения или более явных социальных реакциях, таких как «доминирование — подчинение» в межличностных отношениях. Дальнейшее развитие этого направления в тестировании привело к созданию тестов для количественной оценки выраженности аттитюдов и интересов (глава 14), которые в техническом отношении, по существу, оставались опросниками.

Другой подход к измерению личности представлен применением *тестов действия* (*performance tests*), или *ситуационных тестов* (*situational tests*) (глава 16). В этих тестах от испытуемого требуют выполнить задачу, цель которой часто маскируется. Большинство таких тестов довольно точно моделируют обыденные ситуации. Впервые подобная методика была широко применена в тестах, разработанных X. Хартшорном, М. Мэем и их сотрудниками (Hartshorne, May et al., 1928, 1929, 1930) в конце 1920-х — начале 1930-х гг. Эта серия тестов, стандартизованных на школьниках, имела отношение к таким особенностям поведения, как жульничество, ложь, воровство, действие заодно с товарищами и стойкость. Количественные показатели могли быть получены по каждому из большого набора конкретных тестов. Другой иллюстрацией этого подхода может служить серия ситуационных тестов для взрослых, разработанная в годы Второй мировой войны в рамках аттестационной программы Управления стратегических служб (*OSS*, 1948). Эти тесты предназначались для оценки достаточно сложного и тонкого социального и эмоционального поведения и требовали довольно сложного оборудования и обученного персонала, а способы интерпретации реакций испытуемого оставляли место для субъективности.

Третий подход к изучению личности представлен применением *проективных методик* (глава 15), получивших, особенно у клиницистов, чрезвычайно широкое распространение. В таких тестах клиенту дается неструктурированное задание, предоставляющее широкую свободу в его выполнении. Эти методики основаны на предположении, что в своем решении индивидуум проявит характерные именно для него способы реакции на ситуацию. Подобно тестам действия и ситуационным тестам, проективные методики в большей или меньшей степени маскируют цель обследования и тем самым уменьшают шансы тестируемого человека намеренно создать желаемое впечатление. Уже упоминавшийся тест свободных ассоциаций — один из наиболее ранних типов проективных методик. К этому же типу можно отнести тесты завершения предложений[1]. К заданиям иного типа, обычно применяемым в проективных тестах,

[1] В отечественной литературе этот тип проективных методик часто называют тестами незаконченных предложений. — *Примеч. науч. ред.*

относятся рисование, представляющая сценку расстановка игрушек, импровизация драматической сцены и интерпретация картинок или чернильных пятен.

Применение любых из доступных на данный момент тестов личности связано с серьезными трудностями, как практическими, так и теоретическими. Каждый подход имеет свои преимущества и свои недостатки. В целом же, тестирование личности по всем практическим меркам сильно отстает от тестирования способностей, хотя это отставание не следует приписывать недостатку усилий со стороны ученых. За время, прошедшее с 1950 г., исследования по измерению свойств личности достигли впечатляющего размаха и принесли с собой множество хитроумных приемов и технических усовершенствований в области методов. Медленный прогресс в этой области объясняется скорее особыми трудностями, с которыми сталкивается измерение свойств личности.

В современных исследованиях с использованием тестов личности выявляются две важные объединяющие тенденции (см. Anastasi, 1985b, 1992a, 1993; Digman, 1990; L. R. Goldberg, 1993; Simon, 1994). Во-первых, накапливается все больше данных о взаимовлиянии аффективных («личности») и когнитивных («способностей») свойств человека, причем как при выполнении тестовой задачи, так и в реальном поведении. Традиционное разграничение этих двух типов свойств, или черт, начинают признавать искусственным, принятым в целях удобства при описании и измерении разных сторон поведения. Во-вторых, теоретический анализ природы и структуры личности способствует реинтеграции когнитивных и аффективных свойств в комплексную модель человеческой активности, охватывающую все формы поведения. Эта широкая модель имеет отношение к основным исследованиям как интеллектуальных (глава 11), так и аффективных (глава 13) особенностей.

Часть 2

ТЕХНИЧЕСКИЕ И МЕТОДОЛОГИЧЕСКИЕ ПРИНЦИПЫ



3 НОРМЫ И СМЫСЛОВОЕ ЗНАЧЕНИЕ ТЕСТОВЫХ ПОКАЗАТЕЛЕЙ

Вторая часть учебника, включающая главы 3–7, знакомит с основными понятиями и методологией, необходимыми для понимания психологических тестов и правильной интерпретации их результатов. Соответственно порядку глав в ней рассмотрены нормы, надежность, валидность, анализ заданий и конструирование тестов. Данная глава посвящена разработке и использованию норм, а также другим процедурам, облегчающим пользователям интерпретацию тестовых показателей. При отсутствии дополнительных интерпретирующих данных первичная оценка по любому психологическому тесту лишена всякого смысла. Сказать, что кто-то верно решил 15 задач в тесте математического рассуждения, правильно опознал 34 слова в словарном тесте или успешно собрал механическую конструкцию за 57 с в тесте технических способностей — значит ничего или почти ничего не сообщить о том, как у этого человека развиты соответствующие функции. Знакомые всем процентные показатели также не дают удовлетворительного решения проблемы интерпретации первичных тестовых оценок. Например, 65 % правильных ответов по одному словарному тесту могут означать то же, что 30 % по другому или 80 % по третьему. Разумеется, процентное выражение показателя может иметь тот или иной смысл в зависимости от трудности заданий, из которых состоит каждый тест. Подобно всем первичным оценкам, процентные показатели могут быть истолкованы только в рамках четко заданной и единой системы отсчета.

Оценки по психологическим тестам чаще всего интерпретируются посредством их сопоставления с *нормами*, отображающими выполнение теста в выборке стандартизации. Такие нормы устанавливаются эмпирически, путем определения того, как представители репрезентативной группы в действительности справляются с тестом. После чего первичную оценку («сырой» балл) конкретного человека можно соотнести с распределением оценок, полученных на выборке стандартизации, чтобы узнать, какое место он занимает в этом распределении. Соответствует ли его показатель среднему результату группы, на которой проводилась стандартизация теста? Или же он несколько ниже среднего? А может быть, он попадает в верхний конец распределения и, таким образом, намного превосходит средний результат?

Чтобы более точно определить положение индивидуума относительно выборки стандартизации, его «сырой» балл (первичная оценка) переводится в некую относи-

тельную меру. Предполагается, что эти производные оценки должны служить двум целям. Во-первых, они указывают относительное положение обследованного человека в нормативной выборке и позволяют оценить полученный им результат в сравнении с результатами других людей. Во-вторых, они обеспечивают сопоставимые меры, допускающие прямое сравнение выполнения индивидуумом различных тестов. Например, если девочка получила 40 баллов по словарному тесту и 22 балла по тесту арифметического рассуждения, то это ничего не говорит нам о ее относительной результативности по этим двум тестам. Какой тест она выполнила лучше — словарный или арифметический — или оба одинаково хорошо? Поскольку первичные оценки по разным тестам обычно выражаются в разных единицах, прямое сравнение таких оценок невозможно. Различие в степени трудности еще больше усложняет сравнение первичных оценок по соответствующим тестам. Производные же оценки могут быть выражены в одних и тех же единицах и относиться к одним и тем же или весьма сходным нормативным выборкам для различных тестов. Таким образом, оказывается возможным сравнение относительной эффективности индивидуума при выполнении им множества разных функций.

Есть различные способы преобразования первичных оценок, с тем чтобы они могли служить двум сформулированным выше целям. Однако, с принципиальной точки зрения, получаемые в результате производные оценки выражают один из двух основных аспектов: 1) достигнутый уровень развития или 2) относительное положение индивидуума в определенной группе. Оба типа оценок и некоторые из их распространенных вариантов будут рассмотрены в специальных разделах этой главы. Но прежде необходимо разобраться с несколькими статистическими понятиями, лежащими в основе разработки и использования норм. Цель следующего раздела — разъяснить смысл традиционных статистических мер. Упрощенные вычислительные примеры приведены в нем лишь для иллюстрации и не предназначены для обучения статистическим методам. С формальной стороной вычислений и конкретными алгоритмами решения прикладных задач читатель может ознакомиться по любому современному учебнику статистики для психологов (см., напр.: D. C. Howell, 1997; Runyon, & Haber, 1991; West, 1991). В настоящее время отмечается растущее осознание потребности в элементарных знаниях статистической методологии, причем это касается не только пользователей тестов, но и всех тех, кто хочет с пониманием читать публикуемые материалы исследований в любой области психологии (L. S. Aiken, West, Sechrest,, & Reno, 1990; Anastasi, 1991; Lambert, 1991; S. T. Meier, 1993).

Статистические понятия

Главная цель статистического метода — представить количественные данные в систематизированной и сжатой форме с тем, чтобы облегчить их понимание. Колонка из 1000 тестовых оценок может выглядеть весьма внушительно, но в таком виде она мало что говорит. В качестве первого шага при наведении порядка в этом хаосе «сырых» баллов можно составить таблицу их *частотного распределения* (см. табл. 3–1). Для этого сначала определяются — исходя из числовых значений первичных оценок — удобные интервалы группирования, а затем каждая из этих оценок отмечается условным значком (палочкой, крестиком и т. п.) в соответствующем ей интервале. Когда все первичные оценки разнесены по интервалам группирования, в них подсчитывает-

ся количество условных значков, с тем чтобы найти частоту, или число случаев, для каждого интервала. Сумма всех частот равняется N — общему числу случаев в данной группе. В табл. 3–1 приведены первичные оценки 1000 студентов по тесту усвоения кода, в котором нужно было перейти от использования искусственных слов или бессмысленных слогов из одного набора к пользованию аналогичными элементами из другого набора. Первичные оценки, представленные числом правильных элементов слогового кода, замененных в течение двухминутной попытки, колеблются в пределах от 8 до 52. Они были разнесены по интервалам группирования с шириной 4 единицы: от 8–11 до 52–55. Из колонки частот видно, что оценки двух испытуемых находятся в интервале 8–11, трех — в интервале 12–15, и т. д.

Таблица 3–1

Частотное распределение первичных оценок студентов по тесту усвоения кода (N=1000)

Интервал группирования	Частота
52–55	1
48–51	1
44–47	20
40–43	73
36–39	156
32–35	328
28–31	244
24–27	136
20–23	28
16–19	8
12–15	3
8–11	2

(Из Anastasi, 1934, p. 34)

Информация, содержащаяся в частотном распределении, может быть также представлена графически в виде кривой распределения. На рис. 3–1 данные из табл. 3–1 отображены в графической форме. По горизонтальной оси отложены первичные оценки, представленные границами интервалов группирования, а по вертикальной — частоты, или число случаев, попадающих в каждый интервал. Это график построен двумя способами, в виде гистограммы и полигона (частот), оба из которых достаточно распространены. В *гистограмме* высота столбца над каждым интервалом группирования соответствует числу испытуемых, попавших по результатам тестирования в соответствующий интервал. В *полигоне* число испытуемых в каждом интервале группирования указывается точкой, расположенной над серединой интервала на высоте, соответствующей его частоте, а сами точки последовательно соединяются отрезками прямой.

Если не обращать внимание на некоторые нерегулярности, распределение, представленное на рис. 3–1, имеет сходство с колоколообразной *нормальной кривой.* Математически определенная нормальная кривая изображена на рис. 3–2. Этот тип кривой обладает важными математическими свойствами и лежит в основе многих видов статистического анализа. Для наших целей, однако, достаточно будет отметить лишь некоторые из свойств нормальной кривой. Легко заметить, что согласно нормальному

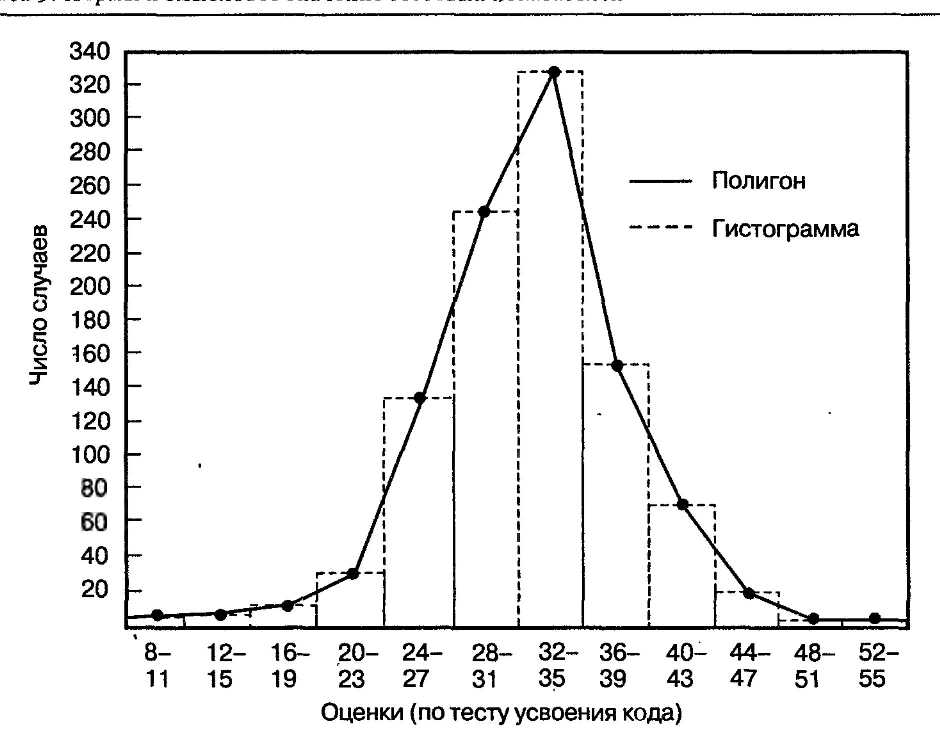


Рис. 3–1. Кривые распределения: полигон и гистограмма (по данным табл. 3–1)

закону распределения наибольшее число случаев скапливается вокруг центральной точки кривой и постепенно падает к ее краям. Кривая симметрична и имеет единственный максимум в центре. Большинство распределений человеческих признаков — от роста и веса до способностей и свойств личности — приближаются к нормальной кривой. В общем, чем больше группа, тем ближе эмпирическое распределение к теоретической нормальной кривой.

Далее, совокупность тестовых оценок может быть сжато описана некоторой мерой *центральной тенденции.* Такая мера дает единственную, наиболее типичную или репрезентативную оценку, характеризующую выполнение теста группой испытуемых, взятой в целом. Самой известной из таких мер является выборочное *среднее* или, точнее, *среднее арифметическое,* обозначаемое чаще всего большой буквой *M* (по первой букве англ. слова *mean*). Оно находится сложением всех оценок и делением получившейся суммы на число случаев (*N*). Другой мерой центральной тенденции является *мода,* или наиболее часто встречающаяся оценка. В частотном распределении мода определяется как середина интервала группирования с максимальной частотой. Например, в табл. 3–1 мода представлена средней точкой интервала 32–35 и равна 33,5. Отметим, что эта величина соответствует самой высокой точке кривой распределения на рис. 3–1. Третья мера центральной тенденции — это *медиана,* или оценка, приходящаяся на середину совокупности ранжированных (упорядоченных по величине) оценок испытуемых. Медиана есть точка, делящая построенное на такой ранжированной совокупности распределение ровно пополам, в результате чего одна половина случаев лежит выше, а другая ниже медианы.

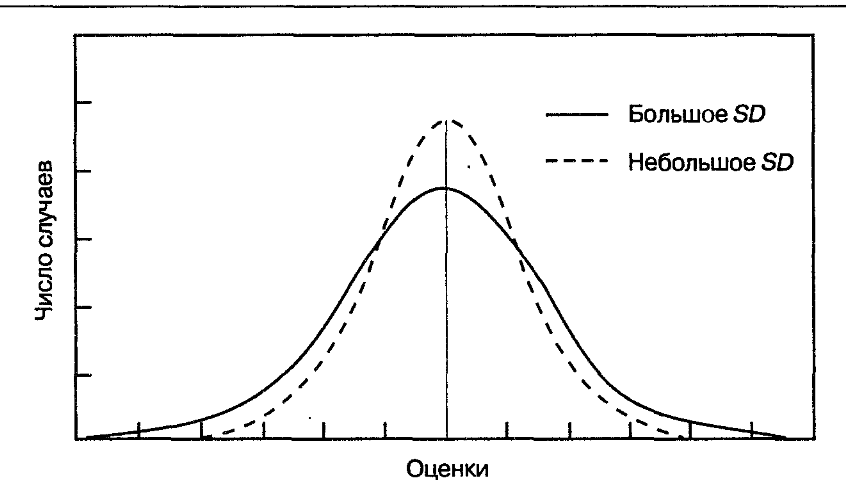


Рис. 3–2. Частотные распределения с одинаковым средним и разным диапазоном изменчивости

Дополнительную информацию о совокупности тестовых оценок дают меры *изменчивости*, показывающие степень индивидуальных отклонений от центральной тенденции. Наиболее очевидным и понятным способом представления изменчивости служит *размах*, определяемый, в простейшем случае, как разность между максимальной и минимальной оценками в совокупности. Однако размах является крайне грубой и неустойчивой мерой изменчивости, поскольку определяется только по двум оценкам. Всего один необычно высокий или низкий результат может заметно повлиять на величину размаха. Более точный метод измерения изменчивости основан на учете разностей между оценками каждого испытуемого и среднегрупповой оценкой.

В этом месте полезно обратиться к примеру в табл. 3–2, где приведены расчеты рассматриваемых нами различных мер для совокупности из 10 случаев. Столь малая совокупность взята для того, чтобы сделать наш пример предельно понятным за счет упрощения вычислений, хотя на практике обычно приходится иметь дело с гораздо большими совокупностями данных. В табл. 3–2 также вводится ряд принятых в статистике обозначений, которые будут использоваться и в дальнейшем. Первичные оценки по тесту по традиции обозначаются прописной буквой X, а строчная буква x служит для обозначения отклонений каждой индивидуальной оценки от группового среднего. Греческая прописная буква Σ расшифровывается как сумма. Среднее значение и медиана вычислены по данным, представленным в первой колонке табл. 3–2. Среднее равно 40; медиана равна 40,5 и находится посередине между оценками 40 и 41: пять случаев (50 %) лежат выше и пять ниже медианы. Находить моду для столь малой совокупности лишено всякого смысла, так как составляющие ее случаи не обнаруживают явного скопления вокруг какой-либо из оценок. Формально, однако, мода представлена оценкой 41, поскольку такую оценку получили два человека, тогда как все другие оценки встречаются лишь по одному разу.

Вторая колонка таблицы показывает, насколько каждая оценка отклоняется в ту или другую сторону от среднегрупповой (40). Сумма этих отклонений всегда равна нулю, так как положительные и отрицательные отклонения от среднего обязательно уравновешивают друг друга (+20 – 20 = 0). Отбросив знаки отклонений и усредняя

Таблица 3–2

Иллюстрация понятий центральной тенденции и изменчивости

	Оценка (X)	Отклонение ($x = X - M$)		Квадрат отклонения (x^2)
	48	+8		64
	47	+7		49
50 % случаев	43	+3	+20	9
	41	+1		1
	41	+1		1
Медиана = 40,5 →				
	40	0		0
	38	−2		4
50 % случаев	36	−4	−20	16
	34	−6		36
	32	−8		64
	$\sum X = 400$			$\sum x^2 = 244$

$$M = \frac{\sum X}{N} = \frac{400}{10} = 40$$

$$\textit{Дисперсия} = \sigma^2 = \frac{\sum x^2}{N} = \frac{244}{10} = 24,40$$

$$SD \textit{ или } \sigma = \sqrt{\frac{\sum x^2}{N}} = \sqrt{24,40} \approx 4,9$$

Примечание. Символы Σ и σ в этой таблице — соответственно прописная и строчная греческие буквы «сигма». Во многих статистических работах символом *SD* (или просто *S*) обозначается выборочное стандартное отклонение, вычисляемое на основе фактически полученных данных, тогда как символ σ используется для обозначения (ожидаемой величины) стандартного отклонения совокупности, из которой извлекалась выборка для сбора данных.

их абсолютные значения, мы можем получить меру средней величины, на которую каждый человек отклоняется от центральной тенденции группы (выраженной средним арифметическим). Несмотря на некоторые достоинства (прежде всего, ясность и понятность) такой дескриптивной меры, «среднее отклонение» не пригодно для более сложного математического анализа данных из-за произвольного отбрасывания знаков и практически не используется в наше время.

Гораздо более полезной мерой изменчивости является *стандартное отклонение* (*SD* или σ), при вычислении которого отрицательные знаки отклонений устраняются математически допустимым способом — путем возведения каждого отклонения в квадрат, как показано в третьей колонке табл. 3–2. Сумма значений в этой колонке, деленная на число случаев $\left(\frac{\sum x^2}{N}\right)$, называется *дисперсией*, или *средним квадратом отклонений*. Дисперсия оказалась крайне полезной при выяснении вкладов разных факторов в индивидуальные различия результатов тестирования. Однако в данный мо-

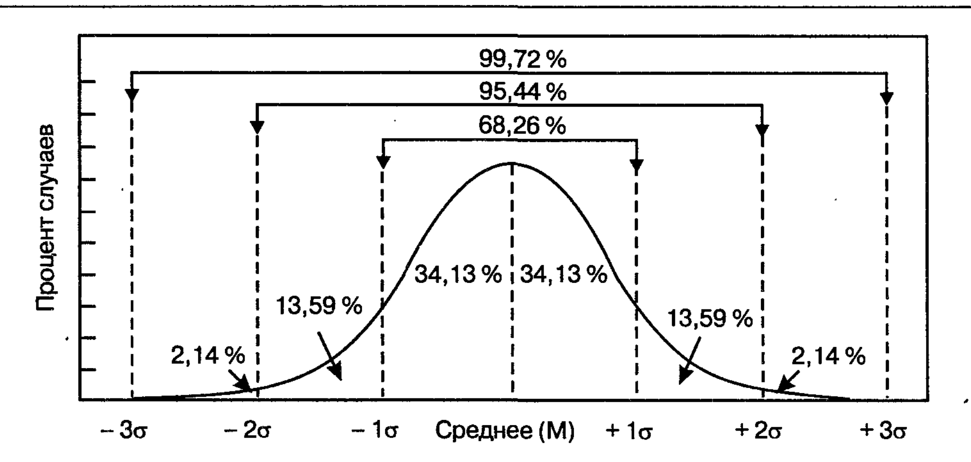


Рис. 3–3. Процентное распределение случаев под нормальной кривой

мент главный интерес для нас представляет стандартное отклонение (*SD*)[1], равное корню квадратному из дисперсии, как видно из табл. 3–2. Эта мера широко используется при сравнении изменчивости данных, полученных в разных группах. На рис. 3–2, например, показаны два распределения с одинаковым средним, но разным диапазоном изменчивости. Распределение с более широким диапазоном индивидуальных различий дает большую величину *SD*, чем распределение с менее выраженными индивидуальными различиями. При оценивании относительных результатов тестирования двух групп мы должны сравнивать не только средние, но и стандартные отклонения. Если эти группы различаются по диапазону изменчивости оценок, это может указывать на различия в доле высоких, низких или тех и других оценок, независимо от различия средних. Современная статистика располагает комплексными методами анализа эффектов, вызванных различиями средних и стандартных отклонений (см., например, Feingold, 1955).

Как будет показано в разделе о стандартных показателях, *SD* также выполняет функцию базисного элемента для выражения оценок индивидуума по различным тестам в единицах норм. Интерпретация стандартного отклонения становится особенно ясной в тех случаях, когда речь идет о нормальной или приблизительно нормальной кривой распределения. При нормальном распределении имеется точное соотношение между *SD* и относительным количеством случаев, как хорошо видно на рис. 3–3. Базис нормальной кривой (ось абсцисс) размечен отрезками, представляющими одно, два и три стандартных отклонения выше и ниже среднего *М*. Например, для данных, приведенных в табл. 3–2, $M = 40 + 1\sigma = 44{,}9$ (т. е. $40 + 4{,}9$); $+ 2\sigma = 49{,}8$ (т. е. $40 + 2 \times 4{,}9$) и т. д. Процент случаев, попадающих в интервал между *М* и $+ 1\sigma$, для нормального распределения равен 34,13 %. Поскольку кривая симметрична, 34,13 %

[1] Иллюстрируемые в этой главе вычисления относятся к *описательной статистике*, применяемой к фактически обследованной выборке; в *статистике вывода* *N* заменяется на *N–1* для того, чтобы получить оценку соответствующих параметров совокупности по выборочным данным. Чем меньше выборка, тем больше будут различия между параметрами генеральной совокупности и их выборочными оценками. За разъяснениями можно обратиться к любому современному учебнику статистики (например, Comrey & Lee, 1992).

случаев попадает также в интервал между М и – 1 σ, так что диапазон от – 1σ + 1σ хватывает 68,26 % случаев. Почти все случаи (99,72 %) лежат в пределах ± 3σ от среднего (*М*). Эти соотношения имеют особое значение для интерпретации обсуждаемых чуть позднее стандартных показателей и процентилей.

Возрастные нормы

Один из способов придать смысл тестовым оценкам — это указать, как далеко продвинулся индивидуум по нормальной траектории развития. Так, можно сказать, что 8-летний ребенок, справляющийся с заданиями теста интеллекта на уровне среднего 10-летнего ребенка, имеет умственный возраст (УВ) 10 лет. Умственно отсталый взрослый, выполняющий задания этого теста на том же уровне, будет также иметь УВ = 10 лет. В другом контексте четвероклассника, например, можно охарактеризовать как достигшего нормы 6-го класса по тесту чтения и нормы 3-го класса по арифметическому тесту. В некоторых системах для описания возрастного развития используются более качественные характеристики изменения специфических функций, таких как сенсомоторная активность или формирование понятий. Но независимо от способа выражения, показатели, основанные на возрастных нормах, довольно грубы и плохо поддаются точной статистической обработке. Тем не менее они имеют сильную притягательность в силу своей наглядности и широко используются, особенно при клиническом обследовании, а также при решении ряда научных проблем.

Умственный возраст. Как отмечалось в главе 2, термин «умственный возраст» получил широкое распространение благодаря различным переводам и адаптациям шкал Бине—Симона, хотя сам Бине пользовался более нейтральным термином «умственный уровень». В таких возрастных шкалах, как шкалы Бине и их последующие редакции (до 1986 г.), тестовые задания группируются по возрастным уровням. Например, задания, посильные для большинства 7-летних детей в выборке стандартизации, относятся к уровню 7 лет; задания, выполняемые большинством 8-летних детей, — к уровню 8 лет и т. д. Казалось бы, в этом случае показатель ребенка по данному тесту должен соответствовать самому высокому возрастному уровню, который ему удалось успешно пройти. В действительности, однако, индивидуальные результаты выполнения теста всегда обнаруживают известную степень *разброса*. Иными словами, обследуемый может не справиться с некоторыми тестами ниже его умственного возраста и выполнить задания, рассчитанные на более высокий умственный возраст. По этой причине сложилась практика, когда сначала определялся *базисный возраст* обследуемого, т. е. максимальный возрастной уровень, на котором и ниже которого все тесты оказываются доступными ребенку. А за все тесты, пройденные на более высоких возрастных уровнях, производились «частичные зачеты» — в месяцах, добавляемых к базисному возрасту. В этом случае умственный возраст ребенка по такому тесту представлял собой сумму базисного возраста и дополнительных «зачетных месяцев».

Нормы в форме умственного возраста использовались и при работе с тестами, которые не подразделялись на возрастные уровни. В таком случае сначала определяется первичная оценка ребенка по тесту (так называемый «сырой» балл). В качестве первичной оценки может выступать просто суммарное количество правильно выполненных заданий всего теста, либо она может быть более сложной и строиться с учетом

времени выполнения заданий, числа ошибок или даже какой-то комбинации таких мер. Средние величины первичных оценок, полученных детьми в каждой возрастной группе выборки стандартизации, и составляют возрастные нормы для такого теста. Например, средняя первичная оценка 8-летних детей могла бы служить нормой для возраста 8 лет. Если первичная оценка обследуемого равна средней первичной оценке 8-летних детей, то его УВ по данному тесту составляет 8 лет. Все первичные оценки по такому тесту можно преобразовать аналогичным способом, соотнося их с возрастными нормами.

Следует отметить, что единица умственного возраста не остается постоянной и с годами обнаруживает тенденцию к сокращению. Так, ребенок, отстающий в развитии на один год в 4-летнем возрасте, к 12 годам будет отставать примерно на 3 года, т. е. один год умственного роста между 3 и 4 годами равносилен 3 годам роста между 9-м и 12-м годом жизни. Поскольку развитие интеллекта идет быстрее в более ранние годы и постепенно замедляется по мере взросления ребенка, единица УВ соответственно уменьшается. Это соотношение можно сделать более наглядным, если представить себе, что рост ребенка выражается в единицах «ростового возраста» (*height age*). Разница, в дюймах, между ростовым возрастом 3 и 4 года будет большей, чем между ростовым возрастом 10 и 11 лет. В силу постепенного сокращения единицы УВ один год опережения или задержки развития в возрасте, скажем, 5 лет означает большее отклонение от нормы, чем тот же год в возрасте 10 лет.

Эквивалентные классы. Показатели тестов достижений в обучении часто интерпретируются в единицах эквивалентных классов. Эта практика вполне понятна, поскольку эти тесты применяются в школьной обстановке. Характеризовать достижения ученика как соответствующие уровню 7-го класса по орфографии, уровню 8-го класса по чтению и уровню 5-го класса по арифметике, для большинства столь же притягательно, как пользоваться понятием умственного возраста в традиционных тестах интеллекта.

Нормы в виде эквивалентных классов определяются посредством вычисления среднего по первичным оценкам, полученным детьми в каждом классе. Так, если среднее количество правильно решенных задач арифметического теста в выборке стандартизации четвероклассников равно 23, то первичная оценка 23 соответствует эквивалентному 4-му классу. Промежуточные эквивалентные классы, представляющие как бы доли класса, обычно определяются путем интерполяции, хотя их можно получить и непосредственно, тестируя детей несколько раз в учебном году. Поскольку учебный год длится 10 месяцев, их последовательность можно представить в виде шкалы десятых долей эквивалентного класса. Тогда 4,0 будет указывать на средний результат выполнения теста в начале обучения в 4-м классе (сентябрьское тестирование), а 4,5 — на средний результат по тому же тесту в середине обучения (февральское тестирование), и т. д.

Несмотря на их популярность, нормы в виде эквивалентных классов имеют ряд недостатков. Во-первых, содержание обучения меняется от класса к классу. Поэтому такие нормы подходят только для общеобразовательных предметов, обучение которым ведется на всех уровнях, охватываемых данным тестом. Они, как правило, неприменимы в старших классах, где многие предметы изучают только один или два года. Даже если предмет преподается на протяжении всего обучения в школе, его значение

может меняться от класса к классу и, следовательно, скорость его изучения может быть различной. Иными словами, единицы шкалы эквивалентных классов явно не равны друг другу, причем отсутствует определенная закономерность в их изменении для разных предметов.

Кроме того, представленные в виде эквивалентных классов нормы могут приводить к ошибочной интерпретации результатов тестирования, если пользователь теста не принимает в расчет способ их получения. Например, если четвероклассник в шкале эквивалентных классов получил оценку 6,9 по арифметике, то это вовсе *не* означает, что он овладел арифметическими операциями, которым обучают в 6-м классе. Бесспорно, он показал такой результат главным образом благодаря отличному знанию арифметики, которую проходят в 4-м классе. И конечно, нельзя считать, что он уже готов к ее изучению по программе 7-го класса. Наконец, нормы в виде эквивалентных классов иногда ошибочно трактуют как нормативы выполнения теста. Учительница 6-го класса, например, может решить, что все ее ученики должны иметь в тестах достижений результаты, соответствующие или по крайней мере близкие к норме 6-го класса. Разумеется, это ошибочное представление не редкость, когда используются нормы в виде эквивалентных классов. Однако индивидуальные различия в пределах одного класса таковы, что диапазон оценок по тесту достижения будет обязательно перекрывать несколько эквивалентных классов.

Порядковые шкалы. Еще один подход к нормам возрастного развития берет начало в исследованиях по детской психологии. Благодаря эмпирическим наблюдениям за развитием младенцев и дошкольников был накоплен обширный материал, позволяющий описать последовательность типичных возрастных изменений таких функций, как локомоция, сенсорное различение, речевое общение и формирование понятий. В качестве первого из таких исследований можно назвать работу А. Гезелла и его коллег по Йельскому университету (Ames, 1937; Gesell, & Amatruda, 1947; Halverson, 1933; Knobloch, & Pasamanick, 1974). «Таблицы развития» Гезелла (*Gesell Developmental Schedules*) позволяют оценить приблизительный уровень развития в месяцах, которого ребенок достиг в каждой из четырех основных областей поведения, именно: двигательного, речевого, приспособительного и лично-социального поведения. Эти уровни определяются сравнением поведения конкретного ребенка с типичным поведением детей в восьми поворотных точках графика возрастного развития, охватывающего диапазон от 4 недель до 36 месяцев.

Гезелл и его сотрудники особо подчеркивали последовательный характер раннего развития поведения. Они приводили обширные данные, свидетельствующие о единообразии хода развития и организации изменений поведения в четкие последовательности. Например, реакции ребенка на помещенный перед ним небольшой предмет обнаруживают характерную хронологическую последовательность в зрительной фиксации и в движениях руки и пальцев. Попытки захватить предмет всей ладонью предшествуют захвату с помощью большого пальца, противопоставляемого остальным четырем, а он, в свою очередь, сменяется более эффективным пинцетным захватом, когда ребенок зажимает предмет между большим и указательным пальцем. Аналогичные последовательные структуры обнаруживаются также в развитии ходьбы, подъеме по лестнице и в большей части сенсомоторного развития первых лет жизни. Шкалы, разработанные в рамках этого подхода, являются порядковыми в том смысле, что смена ста-

дий развития следует неизменному порядку, причем каждая новая стадия предполагает предварительное усвоение поведения, характерного для предыдущих стадий.[1]

В 1960-х гг. резко возрос интерес к теориям развития швейцарского детского психолога Жана Пиаже (см. Flevell, 1963; Ginsburg, & Opper, 1969; D. R. Green, Ford, & Flamer, 1971). Исследования Ж. Пиаже были сосредоточены на развитии когнитивных процессов от младенчества до старшего подросткового возраста. Его больше интересовало развитие специфических понятий, нежели способностей в широком смысле слова. Примером такого понятия, или схемы, может служить постоянство объекта, благодаря которому ребенок сознает тождественность и непрерывность существования объектов, когда они видны под разными углами или находится вне поля зрения. Другим широко изученным понятием является сохранение, т. е. сознавание того, что то или иное свойство объекта сохраняется неизменным, несмотря на воспринимаемые преобразования объекта, как в случаях, когда одно и то же количество жидкости наливается в сосуды разной формы или когда палочки одинаковой длины по-разному располагаются в пространстве.

Задачи Пиаже широко использовали психологи, изучающие возрастное развитие, а некоторые из его задач были организованы в стандартизованные шкалы, которые будут обсуждаться в главе 9 (Goldschmid, & Bentler, 1968b; Pinard, & Laurendeau, 1964; Užgiris, & Hunt, 1975). В соответствии с подходом Пиаже, эти инструменты являются шкалами порядка, в которых достижение той или иной стадии зависит от успешного прохождения более ранних стадий развития измеряемого понятия. Задания в этих шкалах конструируются таким образом, чтобы выявлять главные аспекты каждой стадии развития; и только затем собираются эмпирические данные о возрасте, в котором обычно достигается каждая стадия. В этом отношении данная процедура отличается от процедур, применяемых при построении возрастных шкал, в которых задания отбираются прежде всего по их способности дифференцировать смежные возрасты. Хотя интерес к вкладам школы Пиаже в диагностику психического развития сохраняется, критический теоретический анализ и многочисленные эмпирические проверки этого подхода высветили как его конструктивность, так и ряд ограничений (Sugarman, 1987).[2]

Подводя итог, можно сказать, что порядковые шкалы предназначены для определения стадии, достигаемой ребенком в развитии специфических функций поведения. Хотя получаемые по ним оценки могут сообщаться в виде указания примерных возрастных уровней, такая форма оценок имеет второстепенное значение по сравнению с качественным описанием типичного поведения обследуемого ребенка. Слово «порядок», входящее в название данного типа шкал, указывает на существование единообразия в развитии, проходящем через последовательные стадии. Поскольку эти шкалы обычно дают информацию о том, что конкретный ребенок способен делать в настоя-

---

[1] Данное значение термина «порядковая шкала» отличается от принятого в статистике, где он обозначает любую шкалу, позволяющую упорядочивать различающиеся объекты (или людей) без знания величины различий между ними. В статистическом смысле шкалы порядка противопоставляются шкалам равных интервалов, имеющим единицы измерения. Порядковые шкалы развития ребенка фактически конструируются по образцу шкалы Гуттмана, или модели симплекса, в которой успешное выполнение заданий на одном уровне автоматически предполагает достижение успеха на всех более низких уровнях (L. Guttman, 1944). Расширение анализа Гуттмана с целью включения в него нелинейных иерархий описано у Bart и Airasian (1974), со специальными ссылками на шкалы Пиаже.

[2] Что касается более подробной оценки пиажетианского подхода, см. главу 9.

щее время (например, взобраться по лестнице без посторонней помощи или понять, что количество жидкости сохраняется неизменным при переливании ее в сосуды разной формы), они обладают теми же существенными признаками, что и предметно-ориентированные тесты (*domain-referenced tests*), обсуждаемые в одном из последующих разделов этой главы.

Внутригрупповые нормы

В наше время почти все стандартизованные тесты предусматривают ту или иную форму внутригрупповых норм (*within-group norms*). При наличии таких норм индивидуальный результат тестирования оценивается исходя из выполнения данного теста в наиболее сопоставимой группе стандартизации, как при сравнении полученной ребенком первичной оценки с первичными оценками детей того же возраста или того же года обучения. Внутригрупповые показатели имеют единый и четко определенный количественный смысл и допускают корректное применение большинства методов статистического анализа.

Процентили. Процентильные показатели выражаются в единицах процента лиц, составляющих выборку стандартизации, результат которых ниже установленной первичной оценки. Например, если 28 % людей решают правильно меньше 15 задач в тесте арифметического рассуждения, то первичная («сырая») оценка 15 соответствует 28-му процентилю (P_{28}). Процентиль показывает относительное положение индивидуума в выборке стандартизации. Процентили можно также рассматривать как ранги в группе из 100, с той лишь разницей, что при ранжировании принято начинать отсчет сверху, т. е. с лучшего члена группы, получающего ранг 1. Напротив, в случае процентилей отсчет ведется снизу, так что чем ниже процентиль, тем хуже позиция индивидуума.

50-й процентиль (P_{50}) соответствует медиане — одной из рассмотренных выше мер центральной тенденции. Процентили выше 50-го представляют результаты выше среднего, а процентили ниже 50-го указывают на низкие результаты. 25-й и 75-й процентили называют также 1-м и 3-м *квартилями* (Q_1 и Q_3), поскольку они отсекают нижнюю и верхнюю четверти распределения. Как и медиана, они служат удобными ориентирами для описания распределения показателей и его сравнения с другими распределениями.

Процентили не следует смешивать с привычными для всех процентными показателями. Последние являются первичными оценками и выражаются в единицах процента правильно выполненных заданий, тогда как процентили — это производные оценки, выражающиеся в единицах процента тестируемых. Первичная оценка ниже любой полученной в выборке стандартизации имела бы процентиль, равный нулю (P_0), тогда как первичная оценка, превышающая любую оценку в выборке стандартизации, получила бы процентиль 100 (P_{100}). Эти процентили, однако, вовсе не означают нулевого или абсолютного результата выполнения теста.

Процентильные показатели обладают рядом достоинств. Их легко рассчитать и понять даже сравнительно неподготовленному человеку. Кроме того, процентили имеют универсальное применение. Они в равной мере используются при работе как с детьми, так и со взрослыми, и подходят к любому типу теста, независимо от того, измеряет ли он способности или свойства личности.

Главный недостаток процентилей связан с неравенством их как единиц измерения, особенно на краях распределения. Если распределение первичных оценок приближается к нормальной кривой, что справедливо для большинства тестовых показателей, то различия между первичными оценками вблизи медианы или центра распределения в процентильном выражении преувеличиваются, тогда как аналогичные различия вблизи краев распределения при переведении их в процентили сильно занижаются. Это искажение расстояний между оценками можно увидеть на рис. 3–4. Напомним, что в нормальной кривой случаи тесно сгруппированы в центре и рассеиваются по мере приближения к краям. Следовательно, каждый данный процент случаев вблизи центра соответствует более короткому отрезку на оси абсцисс, чем тот же процент случаев у краев распределения. На рис. 3–4 это несоответствие интервалов между процентилями хорошо заметно, если сравнить расстояние между P_{40} и P_{50} с расстоянием между P_{10} и P_{20}. Еще более разительно несоответствие интервалов между P_{10} и P_{1}. (В теоретической нормальной кривой нулевой процентиль достигается лишь в бесконечности и поэтому не может быть показан на графике.)

То же соотношение можно увидеть, если посмотреть на положение процентилей, соответствующих равным s-интервалам, отложенным в обе стороны от среднего нормальной кривой. Эти процентили приведены в нижней части рис. 3–4. Мы видим, что разность процентилей между средним и + 1σ равна 34 (84–50), а между + 1σ и + 2σ — всего 14 (98–84).

Очевидно, что процентили показывают относительное положение каждого индивидуума в нормативной выборке, а не величину различия между тестовыми оценками. Но если оценки, выраженные в процентилях, наносить на так называемую линейно-вероятностную масштабную бумагу, то и процентильные показатели могут дать адекватную наглядную картину различий между тестовыми оценками. Линейно-вероятностная бумага разграфлена так, что вертикальные линии отстоят друг от друга так же, как и процентили на нормальной кривой (см. рис. 3–4), тогда как горизонтальные линии следуют через одинаковые интервалы, — или наоборот (как на рис. 3–5).

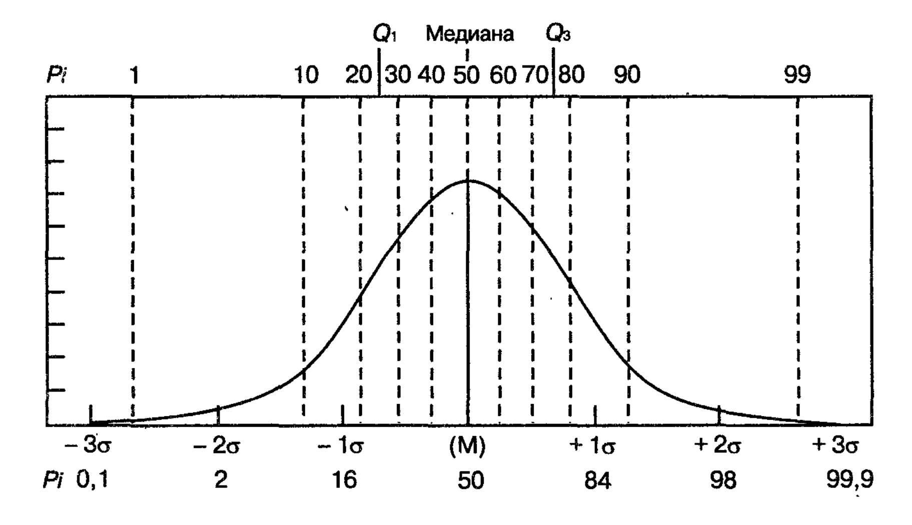


Рис. 3–4. Расположение процентилей при нормальном распределении

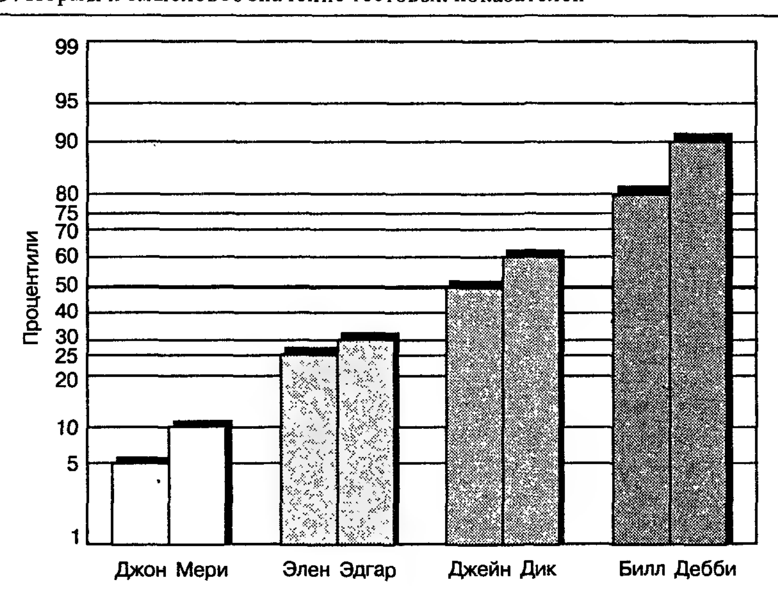


Рис. 3-5. Нормальная процентильная диаграмма. Интервалы между процентилями выбраны таким образом, чтобы соответствовать равным интервалам нормального распределения. Сравним рсстояние между результатами Джона и Мери, с одной стороны, и Элен и Эдгара — с другой: разность процентилей в обоих случаях составляет 5 делений шкалы. В то же время различие между Джейн и Диком, так же как между Биллом и Дебби, составляет 10 делений процентильной шкалы

Такие *нормальные процентильные диаграммы* могут быть использованы для графического представления показателей, полученных разными людьми по одному и тому же тесту, или показателей одного и того же человека по разным тестам. В обоих случаях фактическое различие между показателями будет представлено корректно. Этот способ используется теперь во многих батареях тестов способностей и достижений для вычерчивания профиля оценок, показывающего индивидуальные результаты выполнения каждого теста.

Стандартные показатели. В современных тестах все больше используются стандартные показатели — наиболее удовлетворительный, с точки зрения большинства требований, тип производной оценки. Такие показатели выражают отличие индивидуального результата от среднего в единицах стандартного отклонения соответствующего распределения.

Стандартные показатели могут быть получены как линейным, так и нелинейным преобразованием первичных, «сырых» оценок. При использовании *линейного преобразования* стандартные показатели сохраняют точные численные соотношения первичных оценок, поскольку вычисляются путем вычитания из каждой первичной оценки одной константы и последующим делением разности на другую константу. Относительная величина различий между стандартными показателями, полученными с помощью такого линейного преобразования, в точности соответствует относительной величине различий между первичными оценками. Все свойства исходного распределения «сырых» оценок полностью воспроизводятся в распределении таких стандартных показателей. По этой причине любые вычисления, которые можно производить

Таблица 3–3

Расчет стандартных показателей

$$Z = \frac{X - M}{SD} \qquad M = 60 \qquad SD = 5$$

Результат Элен	Результат Билла
$X_1 = 65$	$X_2 = 58$
$Z_1 = \frac{65 - 60}{5} = +1{,}0$	$Z_2 = \frac{58 - 60}{5} = -0{,}4$

с первичными оценками, можно также выполнять и с линейными стандартными показателями без какого-либо искажения результатов.

Стандартные показатели, получаемые линейным преобразованием, часто называют просто «стандартными показателями» или «*z*-показателями». Чтобы вычислить *z*-показатель, находят разность между первичной оценкой индивидуума и средним для нормативной группы и затем делят эту разность на *SD* нормативной группы. В табл. 3–3 показан расчет *z*-показателей для двух испытуемых, один из которых занимает место на 1 *SD* выше, а другой — на 0,40 *SD* ниже группового среднего. Любая первичная оценка, в точности равная среднему, эквивалентна нулевому значению *z*-показателя. Очевидно, что эта вычислительная процедура будет давать производные оценки с отрицательным знаком для всех лиц с оценками ниже среднего. Кроме того, поскольку для большинства групп область значений индивидуальных оценок не выходит за пределы ± 3 *SD* от среднего, такие стандартные показатели приходится вычислять с точностью хотя бы до десятых, чтобы обеспечить достаточную дифференциацию обследуемых.

Оба этих условия, а именно появление отрицательных величин и десятичных дробей, делают *z*-показатели не слишком удобными для проведения дальнейших вычислений и сообщения результатов. Поэтому обычно применяют еще одно линейное преобразование, единственная цель которого придать показателям более удобную форму. Так, показатели по тестам академической оценки (*SAT*) Совета по вступительным экзаменам в колледжи (*CEEB*) представляют собой преобразованные стандартные показатели со средним *M* = 500 и стандартным отклонением *SD* = 100. Так, стандартный *z*-показатель, равный — 1, в этом тесте выражался бы числом 400 (500 — 100 = 400). Аналогичным образом, *z*-показатель, равный + 1,5, соответствовал бы 650 (500 + 1,5 × 100 = 650). Чтобы перевести стандартный *z*-показатель в эту новую шкалу, нужно просто умножить его на заданную величину *SD*, в данном случае 100, и полученное произведение прибавить (с учетом знака при *z*) к заданному среднему *M* (500). При желании в качестве новых *M* и *SD* можно выбрать любые другие удобные значения; например, показатели по отдельным субтестам в шкалах интеллекта Векслера преобразуются к распределению со средним *M* = 10 и стандартным отклонением *SD* = 3. Все эти меры служат примерами линейно преобразованных стандартных показателей.

Напомним, что одной из причин преобразования первичных оценок в любую производную шкалу выступает стремление добиться сопоставимости показателей по различным тестам. Только что рассмотренные стандартные показатели, получаемые линейным преобразованием, оказываются сопоставимыми лишь в тех случаях, когда распределения «сырых» оценок, по которым они рассчитываются, имеют приблизи-

тельно одинаковую форму. При таких условиях оценка, соответствующая, скажем, + 1 *SD* означает, что индивидуум занимает одинаковое положение относительно обоих групп. Его показатель превышает показатели примерно одинакового процента лиц в обоих распределениях, и этот процент можно определить, когда известна форма распределения. Если же одно распределение заметно скошено, а другое нормально, то *z*-показатель, равный + 1 может превосходить, к примеру, показатели только 50 % членов первой группы и 84 % членов второй.

Чтобы добиться сопоставимости показателей, полученным на основе распределений различной формы, можно применить нелинейное преобразование, позволяющее подогнать показатели к любому заданному типу кривой распределения. Рассмотренные ранее умственный возраст и процентильные показатели представляют собой нелинейные преобразования, но им присущи другие, уже обсуждавшиеся ограничения. Для этой цели обычно используется нормальное распределение, хотя при определенных обстоятельствах другой тип распределения может оказаться более пригодным. Одним из главных доводов в пользу такого выбора является то, что большинство распределений первичных оценок лучше всего аппроксимируется нормальной кривой, чем другими типами кривых. Кроме того, физические характеристики организма, такие как рост и вес, которые измеряются в шкалах с равными единицами, созданных посредством физических операций, обычно имеют нормальное распределение. Другое важное преимущество нормальной кривой заключается в наличии у нее многих полезных математических свойств, облегчающих дальнейшие расчеты.

Нормализованные стандартные показатели — это стандартные показатели, выраженные в единицах распределения, которое было преобразовано с целью его приведения к виду нормальной кривой. Такие показатели можно рассчитывать с помощью таблиц, в которых приводится процент случаев, приходящихся на участки, которые отстоят от среднего нормальной кривой на определенное число единиц *SD*. Сначала определяется процент лиц в выборке стандартизации, приходящихся на (или превышающих) каждую «сырую» оценку. Затем по этому проценту в таблице значений функции плотности нормального распределения отыскивают соответствующее значение нормализованного стандартного показателя. Нормализованные стандартные показатели выражаются в той же форме, что и линейно преобразованные стандартные показатели, т. е. имеют среднее $M = 0$ и стандартное отклонение $SD = 1$. Таким образом, нулевое значение нормализованного показателя показывает, что испытуемый попадает в точку, соответствующую среднему нормальной кривой, превосходя 50 % группы. Показатель, равный — 1, означает, что он превосходит приблизительно 16 % группы, а показатель + 1 — что он превосходит 84 % группы. Эти проценты соответствуют точкам, лежащим соответственно на 1 *SD* ниже и выше среднего нормальной кривой (см. рис. 3–4).

Как и при линейном преобразовании, нормализованным стандартным показателям можно придать любую удобную форму. Например, умножив нормализованный стандартный показатель на 10 и прибавив (по-прежнему с учетом знака) это произведение к 50, получаем *Т-показатель*, предложенный впервые Мак-Коллом (W. A. McCall, 1922). На этой шкале $T = 50$ соответствует среднему, $T = 60$ — превышает среднее на 1 *SD*, и т. д. Еще одно достаточно известное нелинейное преобразование представлено шкалой станайнов, разработанной в ВВС США во время Второй мировой войны. Это шкала одноразрядных оценок со средним $M = 5$ и стандартным отклонением

Таблица 3–4

Значения нормальной плотности (в процентах) для перевода первичных оценок в шкалу станайнов

Процент	4	7	12	17	20	17	12	7	4
Станайн	1	2	3	4	5	6	7	8	9

SD ˜ 2.[1] Название *станайн* (сокращение от англ. *sta*ndard *nine* — стандартная девятка) связано с тем, что оценки в этой шкале принимают значения от 1 до 9.

Первичные оценки можно легко перевести в станайны, упорядочив их по величине и приписав станайны в соответствии со значениями нормальной плотности (в процентах), приведенными в табл. 3–4. Например, если в группе ровно 100 человек, то 4 с самыми низкими первичными оценками получают показатель, равный 1 станайну, следующие 7 — показатель, равный 2 станайнам, следующие 12 — показатель, равный 3 станайнам и т. д. Если группа состоит из большего или меньшего числа обследуемых, то сначала высчитывают, скольким из них соответствует каждый из выписанных в табл. 3–4 процентов, а затем приписывают им соответствующие станайны. Так, при 200 испытуемых 1станайн будет приписан 8 (4 % от 200 = 8), а при 150 — 6 испытуемым (4 % от 150 = 6). Бартлет и Эджертон (Bartlett, & Edgerton, 1966) составили таблицу перевода рангов непосредственно в станайны для групп, содержащих от 10 до 100 человек. Станайны, вследствие их практических и теоретических достоинств, находят все более широкое применение, особенно в тестах способностей и достижений.

Хотя нормализованные стандартные показатели являют собой наиболее удовлетворительный — почти со всех точек зрения — тип показателей, тем не менее имеются определенные технические возражения против нормализации всех распределений подряд. Такое преобразование следует проводить только в тех случаях, когда выборка достаточно велика и репрезентативна и когда есть основания считать, что отклонение эмпирического распределения от нормального произошло в силу определенных недостатков текста, а не особенностей выборки или действия других факторов, влияющих на исследуемое поведение. Следует также отметить, что, когда исходное распределение первичных показателей приближается к нормальному, стандартные показатели, полученные посредством линейного преобразования и нормализации, практически не будут отличаться друг от друга. И хотя методы получения этих двух типов показателей совершенно различны, сами показатели в таких условиях будут почти идентичными. Очевидно, что нормализация распределения, которое и без того фактически нормально, мало или ничего не изменит. Всякий раз, когда это возможно, предпочтительнее добиваться нормального распределения первичных оценок посредством надлежащей коррекции уровня трудности тестовых заданий, а не путем последующей нормализации явно ненормального распределения. В случае приблизительно нормального распределения первичных оценок стандартные показатели, полученные с помощью линейного преобразования, будут служить тем же целям, что и нормализованные стандартные показатели.

---

[1] Кайзер (Kaiser, 1958) предложил модификацию шкалы станайнов, заключающуюся в небольших изменениях процентов и дающую *SD* = 2, что делает ее более удобной в вычислительном отношении. К вариантам этого типа относится *С*-шкала (Guilford & Frucher, 1978, р. 484–487), состоящая из 11 делений и также дающая *SD* = 2, и 10-балльная шкала *стэнов* (сокр. англ. *s*tandard *ten* — стандартная десятка), имеющая по 5 делений в обе стороны от среднего (Canfield, 1951).

Стандартный *IQ* (*deviation IQ*). Для преобразования показателей УВ (умственного возраста) в унифицированный числовой показатель относительного (интеллектуального) статуса индивидуума, в ранних тестах интеллекта был введен коэффициент *IQ* (коэффициент интеллекта). Такой *IQ* определялся просто как отношение умственного возраста (УВ) к хронологическому (ХВ), умноженное на 100 для устранения десятичных дробей (*IQ* = 100 × УВ / ХВ). Очевидно, что если УВ ребенка равен его ХВ, то его *IQ* точно равен 100. *IQ* = 100 означает нормальное или среднее выполнение теста. *IQ* ниже 100 указывает на отставание, а выше 100 — на ускоренное умственное развитие.

Внешняя логическая простота традиционного коэффициента *IQ*, однако, оказалась обманчивой. Главная техническая трудность состоит в том, что, пока стандартное отклонение (*SD*) распределения коэффициентов *IQ* не остается приблизительно постоянным в разных возрастных группах, значения *IQ* у лиц разного возраста будут несопоставимыми. Например, *IQ*, равный 115 в возрасте 10 лет, может указывать на ту же степень превышения среднего уровня, что и *IQ* = 125 для 12 лет, поскольку оба могут приходиться на отметку + 1 *SD* в соответствующих возрастных распределениях. На деле оказалось очень трудно построить тесты, удовлетворяющие психометрическим требованиям сопоставимости коэффициентов *IQ* по всему возрастному диапазону. Главным образом по этой причине простой коэффициент *IQ* сейчас повсеместно заменен так называемым стандартным *IQ*, являющимся по существу еще одной разновидностью уже знакомого стандартного показателя. Стандартный *IQ* представляет собой стандартный показатель со средним 100 и стандартным отклонением, приблизительно равным *SD* распределения *IQ* Стэнфорд—Бине. Хотя стандартное отклонение распределения *IQ* Стэнфорд—Бине (использовалась редакция 1937 г.) не было строго постоянным для всех возрастов, оно колебалось вокруг значения медианы, слегка превышавшего 16. Поэтому если при выборе стандартных показателей для вновь разрабатываемых тестов принять значение *SD*, близкое к 16, то результирующие показатели можно интерпретировать так же, как и *IQ* Стэнфорд—Бине. Поскольку *IQ* Стэнфорд—Бине в ходу уже много лет, тестологи и клиницисты привыкли интерпретировать и классифицировать результаты тестов в единицах уровней такого *IQ*. Они уже знают, чего следует ожидать от лиц с *IQ*, равным 40, 70, 90, 130 и т. д. Таким образом, имеются определенные преимущества в использовании производной шкалы, которая соответствует привычному распределению значений *IQ* Стэнфорд—Бине. Такого соответствия единиц показателей можно достичь подбором численных значений *M* и *SD*, близких к *M* и *SD* распределения *IQ* Стэнфорд—Бине.

Следует добавить, что использование термина «*IQ*» для обозначения таких стандартных показателей может в какой-то степени вводить в заблуждение. Действительно, стандартные *IQ* определяются иначе, нежели традиционные коэффициенты *IQ*. Они *не* являются отношениями умственного и хронологического возраста. И все же употребление применительно к ним традиционного обозначения оправдывается его привычностью, а также тем, что такие показатели *могут* интерпретироваться как *IQ*, при условии приблизительного равенства их *SD* стандартному отклонению ранее известного *IQ*. Среди первых тестов, чьи показатели выражались в единицах стандартного *IQ*, были шкалы интеллекта Векслера со средним *M* = 100 и стандартным отклонением *SD* = 15. Стандартный *IQ* используется в ряде современных групповых тестов интеллекта и в третьей (1960) редакции шкалы интеллекта Стэнфорд—Бине.

В связи с возрастающим применением стандартного *IQ* важно помнить, что стандартные показатели *IQ* из разных тестов сравнимы лишь в тех случаях, когда в их

шкалах используются одинаковые или близкие по величине *SD*. Величину стандартного отклонения следует всегда указывать в руководстве к тесту и учитывать пользователем. Если при построении какой-либо шкалы стандартного *IQ* выбирается иное *SD*, чем в других тестах, то и смысловое значение любого конкретного *IQ* по такому тесту будет существенно отличаться от его смыслового значения в других тестах. Эти расхождения проиллюстрированы в табл. 3–5, где приведены проценты случаев получения показателей *IQ* при нормальных распределениях со стандартными отклонениями от 12 до 18. Эти величины *SD* фактически использованы в шкалах *IQ* ряда опубликованных тестов. Из табл. 3–5 видно, например, что *IQ* ниже 70 отсекает 3,1 % площади под нормальной кривой с *SD* = 16 (как в шкалах Стэнфорд—Бине), но может отсекать всего лишь 0,7 % площади при нормальном распределении с SD = 12 или до 5,1 % при распределении с *SD* = 18. *IQ*, равный 70, традиционно использовался в качестве пограничного значения, отделяющего норму от умственного дефекта. Подобные расхождения, разумеется, имеют место для уровня *IQ* = 130 и выше, который можно использовать при отборе детей для программ работы с интеллектуально одаренными. Диапазон *IQ* = 90–110, обычно характеризуемый как нормальный, может включать от 42 до 59,6 % популяции, в зависимости от выбранного теста. Разумеется, издатели тестов стремятся к унификации, принимая *SD* = 16 в новых тестах и новых редакциях старых тестов, однако сохранившийся разнобой в используемых ныне тестах заставляет каждый раз выяснить величину *SD*.

Соотношения внутригрупповых показателей. На данном этапе рассмотрения производных показателей читатель, вероятно, уже уловил определенную общность между ними. Процентили постепенно приобрели, по крайней мере на графическом уровне, сходство с нормализованными стандартными показателями. Линейные стандартные показатели вообще оказываются неотличимыми от нормализованных, если исходное распределение первичных оценок близко к нормальному. Наконец, стандартные показатели обратились в *IQ*, и наоборот. В связи с последним обстоятельством переосмысление традиционного *IQ*, как в шкале Стэнфорд—Бине, показывает, что эти первые коэффициенты интеллекта (в виде отношения УВ к ХВ) тоже можно интерпретиро-

Таблица 3–5

Процент случаев получения показателей *IQ*, соответствующих разным уровням интеллектуального развития, при нормальных распределениях с *M* = 100 и *SD* = {12, 14, 16, 18}

Уровень *IQ*	Процент случаев							
	SD = 12		*SD* = 14		*SD* = 16		*SD* = 18	
130 и выше	0,7		1,6		3,1		5,1	
120–129	4,3		6,3		7,5		8,5	
110–119	15,2		16,0		15,8		15,4	
100–109	29,8	} 59,6	26,1	} 52,2	23,6	} 47,2	21,0	} 42,0
90–99	29,8		26,1		23,6		21,0	
80–89	15,2		16,0		15,8		15,4	
70–79	4,3		6,3		7,5		8,5	
ниже 70	0,7		1,6		3,1		5,1	
Всего	100,0		100,0		100,0		100,0	

(С любезного согласия Психологической корпорации)

вать как стандартные показатели. Если мы знаем, что распределение коэффициентов *IQ* Стэнфорд—Бине имеет M = 100 и SD ≈ 16, отсюда следует, что *IQ* = 116 превышает среднее на 1 *SD* и совпадает по смыслу со стандартным показателем *z* = + 1,0. Аналогично, *IQ* = 132 соответствует *z* = + 2,0, а *IQ* = 76 эквивалентен *z* = – 1,5 и т. д. Кроме того, показатель *IQ* Стэнфорд—Бине, равный 116, соответствует примерно 84-му процентилю, поскольку 84 % площади под нормальной кривой лежит ниже отметки + 1 SD (рис. 3–4).

На рис. 3–6 показаны соотношения, существующие при нормальном распределении между рассмотренными нами типами показателей, включая *z*-, *T*- и *CEEB*-показатели, стандартный *IQ* Векслера (*SD* = 15), станайны и процентили. Коэффициенты интеллекта (*IQ*) по любому тесту, если они нормально распределены и имеют *SD* = 15, будут совпадать с приведенной здесь шкалой стандартного *IQ*. В эту диаграмму можно

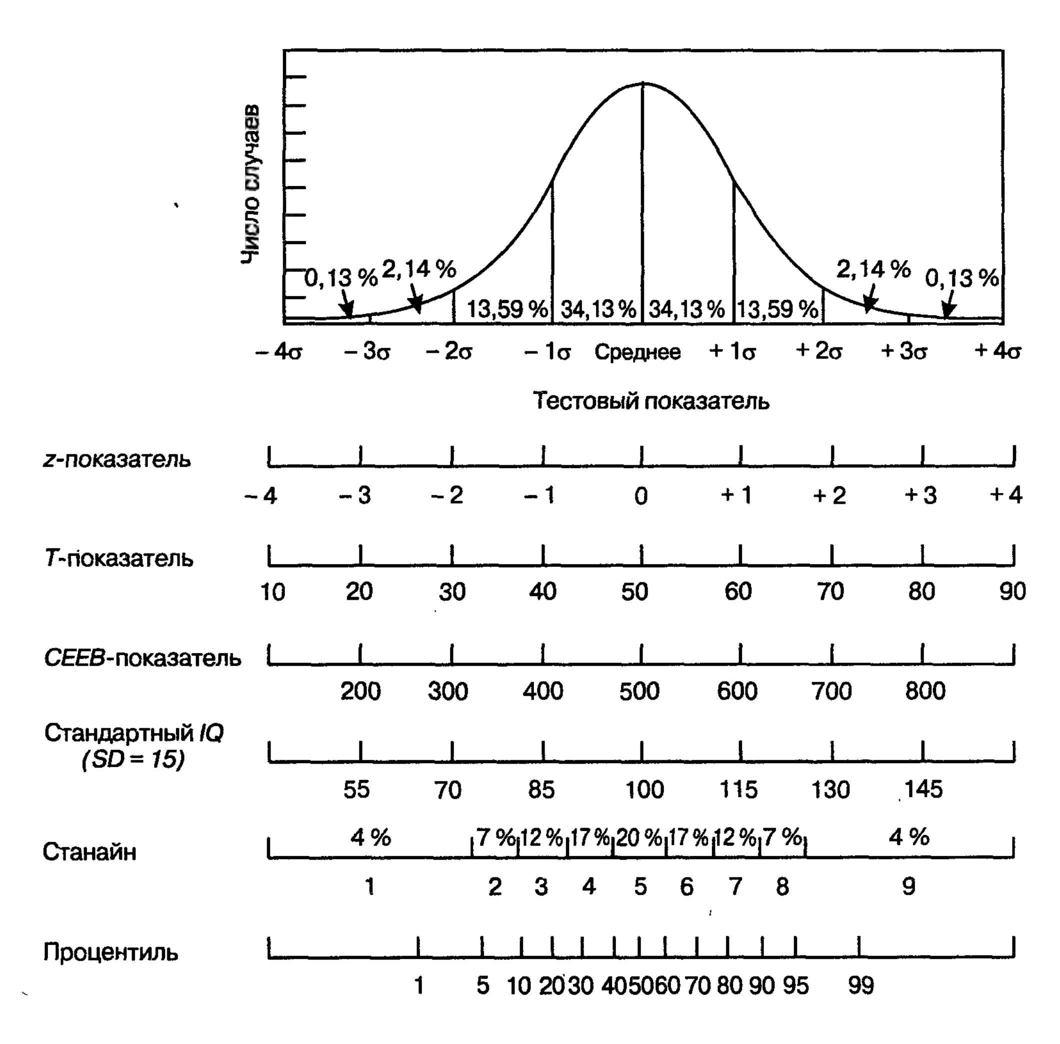


Рис. 3–6. Соотношения между различными типами тестовых показателей при условии нормального распределения

было бы включить любой другой нормально распределенный *IQ* при условии, что мы знаем его стандартное отклонение. Если, например, *SD* = 20, то *IQ* = 120 будет соответствовать + 1 *SD*, а *IQ* = 80, естественно, – 1 *SD*, и т. д.

В заключение отметим, что выбор конкретного вида показателя диктуется главным образом удобством, привычностью и легкостью разработки норм. Ввиду некоторых преимуществ, облегчающих конструирование тестов и статистическую обработку данных тестирования, различные варианты стандартных показателей (включая стандартный *IQ*), в общем, вытеснили остальные типы показателей. Однако большинство типов внутригрупповых производных показателей по существу дела подобны друг другу, если они корректно выводятся и правильно интерпретируются. При соблюдении определенных статистических условий каждый из этих показателей может быть легко переведен в любой другой.

Относительность норм

Межтестовые сравнения. *IQ* или любой другой показатель следует всегда приводить вместе с названием теста, в котором они получены. Тестовые показатели невозможно правильно интерпретировать в отрыве от конкретного теста. Если в школьных записях значится, что Билл Джонс получил *IQ* = 94, а Терри Браун — *IQ* = 110, то эти данные нельзя принимать, так сказать, по нарицательной стоимости без дополнительной информации. Положение этих учащихся вполне могло бы оказаться обратным, доведись им «поменяться» тестами, которые они проходили в своих школах.

Точно так же относительная позиция индивида по различным функциям может быть неверно интерпретирована из-за несопоставимости тестовых норм. Предположим, учащемуся были даны тесты на понимание слов и на способность оперировать пространственными представлениями для оценки его уровня развития в двух соответствующих областях. Если первый из этих двух тестов стандартизован на случайной выборке учеников старших классов, а второй — на специально отобранной группе учеников, посещающих факультативные занятия в школьных мастерских, тестирующий может ошибочно заключить, что этот учащийся гораздо более развит в вербальном, чем пространственном отношении, тогда как на самом деле может иметь место обратное.

Другой пример связан со сравнениями в лонгитюдных исследованиях результатов выполнения теста одним и тем же человеком на разных этапах жизни. Если в личном деле школьника содержатся показатели *IQ*, равные 118, 115 и 101, относящиеся соответственно к 4, 5 и 6-му классам, то первый вопрос, который необходимо задать, прежде чем интерпретировать эти изменения, должен быть таким: «Какие тесты давали в этих трех случаях?» Очевидное снижение результатов может отражать всего-навсего различие между тестами. В этом случае показатели ребенка остались бы теми же, даже если бы эти три теста были проведены с интервалом в одну неделю.

Существуют три основные причины систематических вариаций оценок, получаемых одним и тем же человеком по различным тестам. Во-первых, тесты, даже если они одинаково называются, могут различаться по *содержанию*. Множество примеров тому мы находим среди так называемых тестов интеллекта, обычно фигурирующих под одним и тем же именем, хотя одни из них включают в себя только вербальные задания, другие нацелены преимущественно на проверку пространственных способно-

стей, а третьи могут в равных пропорциях содержать вербальные, пространственные и числовые задания. Во-вторых, иногда несопоставимыми оказываются *единицы измерения* сравниваемых шкал. Как уже объяснялось, если показатели *IQ* по одному тесту имеют *SD* = 12, а по другому — *SD* = 18, то испытуемый, получивший по первому тесту *IQ* = 112, по второму, скорее всего, получит *IQ* = 118. В-третьих, состав *выборок стандартизации,* использованных при определении норм для разных тестов, может оказаться различным. Очевидно, что результаты одного и того же человека будут выглядеть лучше на фоне средних результатов менее способной, чем более способной группы.

Несопоставимость содержания тестов или единиц измерения обычно выявляется при рассмотрении самого теста или при обращении к руководству по его использованию. Но несоответствие нормативных выборок заметить труднее, и им-то, вероятно, и можно объяснить многие не поддающихся иному объяснению расхождения в результатах теста.

Нормативная выборка. Любая норма, как бы она ни выражалась, ограничивается конкретной совокупностью людей, для которой она выводилась. Пользователь теста никогда не должен забывать о том, каким образом устанавливались тестовые нормы. Нормы психологических тестов ни в каком смысле нельзя считать абсолютными, универсальными или постоянными. Они просто отражают уровень выполнения теста лицами, составляющими выборку стандартизации. При формировании такой выборки обычно стремятся получить репрезентативный срез популяции, на которую ориентирован тест.

В статистике принято различать *выборку* и *(генеральную) совокупность.* Первый из этих двух терминов обозначает группу лиц, которые реально проходят тестирование. Второй относится к более широкой, но имеющей тот же состав группе людей, из которой извлекается выборка. Например, если мы хотим установить нормы выполнения теста для совокупности мальчиков 10 лет, живущих в городах и посещающих общественную школу, то нам нужно было бы отобрать, скажем, 500 десятилетних мальчиков, посещающих такие школы в нескольких американских городах. Их выборка, чтобы быть действительно репрезентативной для данной совокупности, должна быть выверена по географическому распределению, социоэкономическому уровню, этническому составу и другим существенным характеристикам.

При разработке и применении тестовых норм на выборку стандартизации следует обращать особое внимание. Очевидно, что выборка, на которой основываются нормы, должна быть достаточно большой для обеспечения их устойчивости. Другая выборка, извлеченная аналогичным способом из той же совокупности, не должна приводить к нормам, заметно отличающимся от полученных. Нормы с большой ошибкой выборки вряд ли добавили бы смысла в интерпретацию тестовых показателей.

Столь же важно, чтобы выборка была репрезентативна изучаемой генеральной совокупности. Необходимо тщательно исследовать даже незначительные факторы, влияющие на отбор испытуемых и делающие выборку нерепрезентативной. Ряд таких факторов можно проиллюстрировать на примере институционных выборок (т. е. выборок из совокупности членов учебных, военных, лечебных, исправительных и других общественных заведений). Использование таких выборок ввиду их доступности и возможности привлечения большого числа испытуемых представляется заманчивым для сбора нормативных данных. Однако нужно внимательно анализировать присущие этим выборкам ограничения. Так, тестированию школьников свойственно постепен-

ное от класса к классу повышение уровня испытуемых, вследствие отсева менее способных учеников. В различных подгруппах это явление выражено неодинаково. Например, процент выбывших выше для мальчиков, чем для девочек. Он также выше для социальных групп, находящихся на более низком экономическом уровне.

Факторы отбора действуют и в других институционных выборках, например в выборках заключенных, пациентов психиатрических больниц или интернатов для умственно отсталых. Благодаря конкретным причинам, определившим помещение индивидуума в специальное учреждение, упомянутые группы не репрезентативны генеральной совокупности преступников, душевнобольных или умственно отсталых. Так, умственно отсталые, страдающие физическими недостатками, чаще оказываются в специальном учреждении, чем физически полноценные. Аналогично этому, доля лиц с глубокой умственной отсталостью будет намного больше в выборке такого типа, чем в соответствующей генеральной совокупности.

С вопросом репрезентативности выборки тесно связана потребность точного определения совокупности, на которую можно распространить полученные нормы. Очевидно, одним из способов обеспечения репрезентативности выборки является ограничение совокупности в соответствии с техническими характеристиками выборки. Например, если генеральная совокупность определяется так, чтобы включать не всех вообще 14-летних детей, а только 14-летних школьников, то при таких ограничениях школьная выборка могла бы быть репрезентативной. В идеале, разумеется, желаемая совокупность должна определяться заранее, исходя из целей теста, а уж затем формироваться выборка. Невозможность привлечь нужных испытуемых может, однако, сделать эту цель недостижимой. В таком случае лучше переопределить более узко изучаемую совокупность, чем распространять нормы на генеральную совокупность, которая не была адекватно представлена выборкой стандартизации. На самом деле лишь очень малое число тестов стандартизовано на таких широких совокупностях, как это обычно представляется непрофессионалам. Тестовых норм, действительных для всего рода человеческого, не существует! Сомнительно также, чтобы по какому-либо тесту имелись адекватные нормы для таких широко определяемых совокупностей, как «взрослые американцы-мужчины», «американские дети 10-летнего возраста» и т. п. Следовательно, выборки, получаемые различными создателями тестов, могут и не представлять в полной мере предполагаемые ими совокупности, обнаруживая смещенность в тех или иных отношениях. Отсюда и несопоставимость получаемых норм.

При интерпретировании тестовых показателей пользователю теста следует принимать во внимание специфические факторы, которые могли повлиять на нормативную выборку, использовавшуюся при стандартизации данного конкретного теста. К ним можно причислить особые факторы отбора, а также господствующие общественные условия в период сбора нормативных данных (Anastasi, 1985d).

Национальные анкерные нормы. Одно из решений проблемы несопоставимости норм заключается в использовании анкерного теста для составления *таблиц эквивалентности* показателей разных тестов. Назначение таких таблиц — представление информации о том, какой показатель в тесте *A* эквивалентен каждому показателю в тесте *B*. Для их построения можно воспользоваться *методом равных процентилей*, согласно которому показатели считаются эквивалентными, если они имеют равные процентили в данной группе. Например, если 80-й процентиль в одной и той же группе соответствует *IQ* = 115 по тесту *A* и *IQ* = 120 по тесту *B*, то *IQ* = 115 в тесте *A* считается эквивалентным *IQ* = 120 в тесте *B*. Этот метод в ограниченной степени практиковался

некоторыми издателями тестов, выпустившими таблицы эквивалентности для нескольких собственных тестов (напр., Lennon, 1966a).

Время от времени делались попытки реализовать более честолюбивые замыслы, в частности откалибровать каждый новый тест относительно единого анкерного теста, который был проведен на высоко репрезентативной нормативной выборке в масштабах всей страны (Lennon, 1966b). Пример — исследовательская программа «Анкерный тест» (*Anchor Test Study*), проведенная Службой тестирования в образовании при поддержке Федерального управления просвещения (*U. S. Office of Education*) (Jaeger, 1973). Ее целью было получение сопоставимых и действительно репрезентативных общенациональных норм по семи наиболее употребительным тестам достижений в области чтения, предназначенным для учеников средних классов. По тщательно разработанному плану эксперимента, позволявшему контролировать многие переменные, в 50 штатах были обследованы свыше 300 000 учеников 4, 5 и 6-х классов. Анкерный тест состоял из субтестов понимания прочитанного и словарного запаса, входящих в Тест достижений для учащихся американских школ (*Metropolitan Achievement Test*), по которым на первом этапе исследования были установлены новые нормы. На этапе калибровки каждому ребенку предъявлялись субтесты понимания прочитанного и словарного запаса двух из семи батарей, причем план эксперимента предполагал использование всех сочетаний из семи батарей по две. Некоторым группам предъявлялись параллельные формы двух субтестов из одной и той же батареи. В специальных группах предъявление всех пар субтестов осуществлялось в обратной последовательности, что позволяло контролировать влияние порядка проведения тестов. По результатам статистического анализа полученных данных были составлены, с помощью метода равных процентилей, таблицы эквивалентности показателей для семи тестов, а также подготовлено руководство по интерпретации их показателей для работников системы образования и других заинтересованных лиц (Loret, Seder, Bianchini, & Vale, 1974).

Впоследствии данные, собранные на калибровочном этапе программы «Анкерный тест», были использованы для разработки шкалы единого показателя, получившей название Национальной эталонной шкалы (*National Reference Scale*) (Rentz, & Bashaw, 1977). Разработанные таким образом таблицы перевода позволяют преобразовать показатель учащихся соответствующих классов по любому из семи тестов (включая их параллельные формы) в трехместный показатель единой непрерывной шкалы. Эта шкала была построена благодаря применению методов анализа заданий и шкалирования, основывающихся на модели Раша; одна из простейших моделей анализа заданий рассматривается позже в этой главе и более полно — в главе 7.

Для многих целей тестирования полезно иметь сопоставимые показатели по разным тестам, которые выражались бы в единицах одной измерительной шкалы и были выверены на одной нормативной выборке. Следует, однако, заметить, что есть разные степени и виды сопоставимости показателей. Сопоставимость, достигаемая в конкретных ситуациях, зависит от сходства тестов по содержанию и таких психометрических свойств, как надежность и уровень трудности, а также от статистических методов, используемых для получения сопоставимых показателей (Angoff, 1984; Angoff, & Cowell, 1986; P. W. Holland, & Rubin, 1982). Не стоит характеризовать тесты как приравненные или полностью эквивалентные, если они не допускают взаимозамены. Несмотря на это, различные виды и степени сопоставимости *могут* облегчить интерпретацию результатов тестирования, при условии, что сравниваемые показатели используются правомерно и с полным представлением о том, как они были получены.

Специфические нормы. Другой, и для большинства тестов, вероятно, более реалистический подход к решению проблемы неэквивалентности существующих норм заключается в стандартизации тестов на более узко определяемых совокупностях, выбираемых сообразно специфическим целям каждого теста. В таких случаях границы нормативной выборки должны быть четко определены и приведены вместе с нормами. Так, о нормах может быть сказано, что они применимы к «конторским служащим крупных фирм» или к «студентам-первокурсникам машиностроительных факультетов университетов». Для многих целей тестирования желательно иметь высоко специализированные нормы. Даже когда имеются репрезентативные нормы для более широко определяемой генеральной совокупности, часто оказывается полезным располагать отдельно публикуемыми *нормами для подгрупп.* Они явно не будут лишними в тех случаях, когда показатели теста заметно меняются от одной группы к другой. Сами подгруппы могут формироваться по признаку возраста, года обучения, типа школьной программы, пола, географического региона, проживания в городе или в сельской местности, социоэкономического уровня и т. д. А предназначением теста будет определятся наиболее существенный признак дифференциации подгрупп, равно как и целесообразность применения общих или специфических норм.

Следует также упомянуть о *локальных нормах,* которые нередко разрабатываются самими пользователями тестов в конкретных социальных условиях. Группы, используемые для получения таких норм, еще более специфичны, чем даже обсуждавшиеся выше подгруппы. Так, работодатель может накапливать нормы, тестируя претендентов на определенные должности в конкретной компании; приемная комиссия колледжа может разрабатывать нормы, обследуя совокупность своих студентов, а какая-то начальная школа может оценивать выполнение тестов своими учениками на основе собственного, внутришкольного распределения показателей. Эти локальные нормы в большей степени, чем общенациональные, отвечают таким задачам тестирования, как предсказание учебных (студенческих) или профессиональных достижений, сравнение относительных успехов детей по различным предметам, измерение .

Фиксированная эталонная группа. Хотя способ вычисления большинства производных показателей предусматривает непосредственную нормативную интерпретацию выполнения теста, существуют и примечательные исключения. Один тип ненормативных шкал использует фиксированную эталонную группу для обеспечения *сопоставимости и преемственности* показателей, не предусматривая нормативного оценивания выполнения теста. При использовании такой шкалы нормативная интерпретация требует обращения к независимо накопленным нормам в ходе обследования подходящей совокупности лиц. Нередко для этой цели используются локальные или другие специфические нормы.

Одним из самых ранних примеров шкалирования в единицах показателей фиксированной эталонной группы служит шкала Теста академических способностей (*Scholastic Aptitude Test* или, сокращенно, *SAT*)[1] Совета колледжей (Donlon, 1984). В период между 1926 г. (когда этот тест был применен впервые) и 1941 г. показатели *SAT* выра-

[1] Позднее этот тест был переименован в *Тест академической оценки* (*Scholastic Assessment Test*) с целью отразить изменение взглядов на природу тестовых показателей, которое произошло в конце XX столетия. (См. особенно главу 12 о влиянии индивидуальных различий жизненного опыта на выполнение теста.)

жались в нормативной шкале, исходя из среднего и *SD* оценок абитуриентов, полученных при очередном проведении теста. По мере того как увеличивалось число и разнообразие колледжей — членов Совета и, соответственно, менялся состав совокупности абитуриентов, было решено сохранить преемственность шкалы *SAT*, ибо в противном случае индивидуальные показатели ставились бы в зависимость от особенностей контингента, проходящего обследование в данном году. Еще более актуальный повод для сохранения преемственности шкалы дало наблюдение, согласно которому учащиеся, проходившие *SAT* в одно время года, справлялись с ним хуже тех, кто проходил тестирование в другое время года, вероятно, вследствие различного действия факторов отбора. Поэтому после 1941 г. все показатели *SAT* стали выражаться в единицах шкалы, в основу которой положено среднее и *SD* оценок примерно 11 000 абитуриентов, проходивших этот тест в 1941 г. Эти абитуриенты и составили фиксированную эталонную группу, используемую при пересчете показателей всех последующих форм данного теста. Например, показатель 500 любой формы *SAT* соответствует среднему в выборке 1941 г.; показатель 600 превышает среднее на 1 *SD*, и т. д.

Для того чтобы можно было перевести первичные показатели любой формы *SAT* в показатели фиксированной эталонной группы, в каждую такую форму включен короткий анкерный тест (или набор общих заданий). Тем самым каждая новая форма связывается с одной или двумя более ранними формами, а те, в свою очередь, — с другими, еще более ранними, цепочкой заданий, доходящей до исходной формы 1941 г. Эти ненормативные показатели *SAT* можно к тому же интерпретировать, сопоставляя с любым подходящим распределением оценок, таким как распределение показателей конкретного колледжа, колледжей определенного типа, региона и т. д. Подобные специфические нормы более полезны для принятия решений о приеме в колледж, чем, скажем, ежегодные нормы, основанные на результатах тестирования полной совокупности абитуриентов. Кроме того, любые происходящие со временем изменения в совокупности абитуриентов можно обнаружить только пользуясь шкалой фиксированных показателей. Совсем недавно шкала *SAT* была заново откалибрована по результатам более миллиона учащихся, закончивших среднюю школу в 1990 г. и прошедших этот тест во время обучения в младшей средней (9–10 кл.) или старшей средней (11–12 кл.) школе. Показатели учащихся, выполняющих *SAT* после 1 апреля 1995 г., заносятся в табель успеваемости уже в единицах шкалы, перестроенной на основе эталонной группы 1990 г. Для пользователей SAT были разработаны разъяснительные материалы и вспомогательные средства для облегчения перевода индивидуальных и совокупных показателей из старой шкалы в новую и наоборот (см. главу 17). Таким образом созданы условия для полной и разнообразной интерпретации индивидуальных результатов в соответствии со специфическими целями тестирования.[1]

Шкалы, построенные по данным фиксированной эталонной группы, в одном отношении аналогичны физическим измерительным шкалам. В этой связи Ангофф (Angoff, 1962, p. 32–33) пишет:

> Вряд ли кто теперь точно знает первоначальное определение длины фута, которым пользуются для измерения высоты и расстояния. Вряд ли кто назовет имя короля, чья ступня была принята в качестве эталона. Вместе с тем мало

---

[1] Мы выражаем благодарность Уэйну Камара из Совета колледжей за помощь в получении сообщаемой здесь информации.

таких, кто не смог бы оценить длину или расстояние с помощью этой единицы измерения. Наше незнание буквального значения или происхождения фута ни в коей мере не делает его бесполезным, ведь, сколько бы ни прошло времени, фут останется одним и тем же, и это позволяет нам освоиться с ним. То же самое можно сказать и про другие единицы измерения — дюйм, милю, градус Фаренгейта и т. д. В области психологического измерения столь же справедливо утверждение, что из первоначального определения шкалы ничего не следует или не должно следовать. Все, что требуется — сохранять постоянной шкалу (в программах тестирования с применением множества форм это достигается их попарным приравниванием) и обеспечивать своевременный приток дополнительных нормативных данных, обновляемых по мере необходимости, которые облегчают интерпретацию и принятие конкретных решений.

Теория «задание — ответ». Семидесятые годы были отмечены всплеском интереса к семейству довольно сложных в математическом отношении процедур для шкалирования тестовых заданий по уровню трудности (Hambleton, 1989; Hambleton, Swaminathan, & Rogers, 1991; Jaeger, 1977). Поскольку эти процедуры требовали большого объема вычислений, их практическое применение стало возможным только с появлением широкого доступа к быстродействующим вычислительным машинам. Существенно различаясь по сложности и используемым математическим методам, все эти подходы первоначально были объединены под общим названием: *модели латентных черт.* В качестве основной меры в них выбиралась вероятность того, что человек с определенной способностью (так называемой латентной чертой) преуспеет в выполнении задания установленной трудности. Однако при этом не подразумевалось, что такие латентные черты или базисные способности существуют к каком-то физическом или физиологическом смысле и что они служат причинами поведения. Латентные черты — всего лишь статистические конструкты, которые математически выводятся из эмпирически измеренных связей между ответами на тест. Грубой, первичной оценкой латентной черты обследуемого является совокупный показатель, полученный им по данному тесту. Во избежание ошибочных мнений, создаваемых термином «латентная черта», некоторые из ведущих представителей этого подхода заменили его более точным описательным термином «теория “задание — ответ”» (*item response theory*) или, сокращенно, IRT (Lord, 1980; D. J. Weiss, & Davison, 1981). И именно это название стало общеупотребительным в психологии.

По существу, *IRT*-модели используются для создания унифицированной — «независимой от выборки» — измерительной шкалы, применимой к отдельным лицам и группам лиц с широко варьирующим уровнем способности и пригодной для широко варьирующего по уровню трудности содержания теста. Как и в случае с фиксированной эталонной группой, описанной в предыдущем разделе, *IRT*-модели требуют анкерных заданий или общего теста в качестве устройства сопряжения между выборками обследуемых и между тестами или наборами заданий теста. Однако, вместо того чтобы использовать для определения нулевой точки и единицы шкалы среднее и *SD* специфической эталонной группы, в *IRT*-моделях эти параметры шкалы устанавливаются на основе данных, представляющих широкий диапазон способности и трудности задания, которые могут собираться на разных выборках. Обычно нулевую точку шкалы устанавливают в центральной области этого диапазона. Единица общей шкалы математически выводится из данных, касающихся заданий; такой подход имеет ряд пре-

имуществ, как теоретических, так и практических, перед более ранними методами анализа заданий. Конкретные аспекты методологии *IRT* обсуждаются в главе 7, в связи с рассмотрением всей совокупности методов анализа заданий. Постепенно *IRT* внедряется в крупномасштабные программы тестирования. Например, начиная с 1982 г., методы *IRT* применяются для приравнивания суммарных показателей по новым формам *SAT*, чтобы выражать их в неизменной, единой шкале (Camara, Freeman, & Everson, 1996; Donlon, 1984).

Общей проблеме *приравнивания тестов (test equating)*, посредством чего показатели по разным формам теста выражаются в показателях единой шкалы, всегда уделялось неослабное внимание. Рассмотрение специальных вопросов различных подходов к этой проблеме выходит за рамки этого учебника. Исчерпывающий обзор и критическую оценку существующих на данный момент методов приравнивания тестов читатель может найти в работах P. W. Holland, & Rubin (1982) и Petersen, Kolen, & Hoover (1989).

Компьютеры и интерпретация тестовых показателей

Технический прогресс. Компьютеры оказали заметное влияние на все этапы тестирования — от конструирования теста до его проведения, подсчета «сырых» баллов, сообщения результатов и их интерпретации (F. B. Baker, 1989; Butcher, 1987; Gutkin, & Wise, 1991; Roid, 1986). Очевидные выгоды от использования компьютеров, даже самых первых, связывают с буквально небывалым увеличением скорости, с какой осуществляется анализ данных и подсчет показателей. Выигрыш от применения компьютеров для автоматизированного проведения традиционных тестов можно отнести к той же категории, поскольку они облегчают и улучшают процедуры проведения таких тестов. Однако гораздо важнее вклад вычислительной техники в разработку новых методов и подходов в психологическом тестировании, которые были бы невозможны без гибкости и мощности, обеспечиваемых современными компьютерами при обработке информации. Иллюстрацией влияния компьютеров в этой области может служить возрастающее применение *IRT*-моделей для создания независимых от выборки шкал, упоминавшихся в предыдущем разделе. Другие новшества в тестировании, явившиеся результатом применения компьютеров, обсуждаются при рассмотрении соответствующих тем на протяжении всей книги.

В связи с темой этой главы мы рассмотрим применение компьютеров для оценки результатов выполнения теста (F. B. Baker, 1989; Gutkin, & Wise, 1991; Roid, & Gorsuch, 1984). На простейшем уровне большинство современных тестов, особенно групповых, теперь приспособлено для *машинного подсчета первичных показателей (computer scoring)*. Некоторые издательства тестов, а также ряд независимых организаций по обработке результатов тестирования, оснащены необходимым оборудованием для предоставления соответствующих услуг пользователям тестов. Кроме того, все более доступными становятся компьютерные диски, с помощью которых пользователи тестов могут обрабатывать результаты тестирования на своих собственных компьютерах (например, программы *ASSIST*, разработанные American Guidance Service). На более сложном уровне доступна *описательная машинная интерпретация (narrative compu-*

ter interpretation) результатов тестирования, правда, лишь для некоторых тестов. В таких случаях специфические паттерны ответов связываются машинной программой с теми или иными словесными формулировками, хранящимися в памяти машины. Этот подход был реализован в отношении как тестов личности, так и тестов способностей. Например, работая с Миннесотским многофазным личностным опросником (*MMPI*), рассмотренным в главе 13, пользователи наряду с числовыми показателями могут получить распечатку диагностических и интерпретационных формулировок о тенденциях личности обследуемого и его эмоциональном состоянии. Для пользователей тестов, имеющих доступ к компьютерам, появляется все больше возможностей приобрести программы, которые выдают не только числовые показатели, но и содержащие их толкование текстовые отчеты по ряду тестов, таких, например, как шкалы интеллекта Векслера для детей (*WISC-R*) и взрослых (*WAIS-R*).

Индивидуализированная интерпретация тестовых показателей на еще более сложном уровне иллюстрируется *интерактивными компьютерными системами (interactive computer systems)*, в которых человек напрямую связан с компьютером через устройства ввода и, фактически, вовлекается в диалог с ним (J. A. Harris, 1973; Holtzman, 1970; M. R. Katz, 1974; Super et al., 1970). Такие диалоговые системы опробовались и изучались в области выбора дальнейшего образования и карьеры, а также на других моделях принятия решения. В подобной ситуации тестовые показатели обычно вводят в компьютерную базу данных наряду с другой информацией, поступающей от учащегося или клиента. По существу, компьютер объединяет всю доступную информацию о конкретном человеке с хранящимися в памяти данными об образовательных программах и профессиях и использует все относящиеся к делу факты и связи, отвечая на вопросы этого человека и помогая ему прийти к какому-то определенному решению. В качестве примера таких интерактивных компьютерных систем можно привести профориентационную диалоговую систему «*SIGI*» (*System for Interactive Guidance Information*, 1974–1975). После десятилетнего периода использования в колледжах и университетах эта система была обновлена и пересмотрена с тем, чтобы отвечать запросам не только студентов, но и зрелых людей, решившихся выйти на рынок труда, сменить профессию или обдумывающих возможности служебного роста (M. R. Katz, 1993; Norris, Schott, Shatkin, & Bennett, 1986).

Опасности и руководящие принципы применения компьютеров в тестировании. Несмотря на то что компьютеры, бесспорно, открыли путь для беспрецедентных усовершенствований всех аспектов психологического тестирования, в некоторых случаях их применение может приводить к неправильному использованию и толкованию тестовых показателей (Butcher, 1985a; J. J. Kramer, & Mitchell, 1985; Matarazzo, 1983, 1986a, 1986b). В связи со стремлением принять соответствующие меры предосторожности значительное внимание было уделено разработке руководящих принципов тестирования с использованием компьютеров. *Стандарты тестирования* (*AERA*, *APA*, *NCME*, 1985) включают ряд стандартов в отношении такого тестирования. Кроме того, был специально разработан комплекс более подробных инструкций в отношении применения компьютеров в различных областях и на разных этапах тестирования (см., например, Butcher, 1987, p. 413–431). Что касается всесторонней оценки использования компьютеров в тестировании, включая машинную интерпретацию показателей, см. Moreland (1985, 1992).

Два из основных вопросов, вызывающих особую озабоченность в связи с распространением компьютерного тестирования, имеют отношение к сопоставимости пока-

зателей и машинной интерпретации результатов теста. В тех случаях, когда один и тот же тест проводят в компьютерной и традиционной бланковой форме, надо проводить специальное исследование сопоставимости показателей (Mazzeo, Druesne, Raffeld, Checketts, & Muhlstein, 1991). Пока не доказано, что эти две формы теста являются полностью эквивалентными, к ним нельзя применять один и тот же набор норм; кроме того, надежность и валидность теста могут варьировать в зависимости от формы предъявления. Особенно важно контролировать сопоставимость показателей у разных людей или групп, чей опыт пользования компьютером, и особенно компьютерного тестирования, может существенно различаться.

Быстрый рост услуг в области машинной интерпретации результатов тестирования с предоставлением готовых отчетов вызвал особую озабоченность. Два основных принципа лежат в основе большинства относящихся к этому вопросу инструкций и руководств. Согласно первому принципу, пользователю теста должна быть предоставлена соответствующая информация, позволяющая оценить надежность, валидность и другие технические характеристики интерпретирующей системы, использованной при разработке программного обеспечения. Каким образом интерпретирующие формулировки выводились из показателей? Какое теоретическое обоснование и эмпирическое подтверждение получила система машинной интерпретации? Основываются ли интерпретирующие формулировки на результатах количественного анализа данных или на суждениях экспертов? Если имеет место последнее, то должны быть представлены сведения о квалификации участвовавших экспертов.

Согласно второму принципу, в тех случаях, когда машинные интерпретации результатов тестирования используют в клинической диагностике, консультировании или в каких-то других областях принятия важных решений в отношении конкретного человека, совершенно необходимо принимать в расчет другие доступные источники информации о тестируемых людях. По этой причине машинными интерпретациями должны пользоваться только высококвалифицированные профессионалы. Такие интерпретации следует рассматривать как средство облегчения работы специалиста, а отнюдь не как его возможную замену.

Интерпретация предметно-ориентированных тестов

Природа и направления использования. Подход к тестированию, вызвавший волну активности, особенно в сфере образования, вначале был назван «критериально-ориентированным тестированием» (*criterion-referenced testing*). Этот термин, впервые предложенный Р. Гласером (R. Glaser, 1963), употребляется до сих пор, причем достаточно вольно, и различные авторы определяют его по-разному. Кроме того, появился ряд альтернативных терминов: содержательно- (*content-*), предметно- (*domain-*) и задачно- (*objective-*) ориентированное тестирование. Они иногда употребляются как синонимы термина «критериально-ориентированное тестирование», а иногда с целью подчеркнуть несколько иные смысловые акценты. Постепенно первоначальное название «критериально-ориентированное тестирование» было вытеснено из обращения более точными описательными терминами. В этой книге предпочтение отдано одному из таких терминов — «предметно-ориентированное тестирование» (*domain-referenced testing*), который и будет употребляться впредь.

Для предметно-ориентированного тестирования типично использовать в качестве интерпретационной системы отсчета не заранее оговоренную совокупность *людей*, а строго определенную *содержательную* область. В этом отношении оно с самого начала противопоставлялось обычному, ориентированному на нормы тестированию, в котором показатель каждого конкретного человека интерпретируется посредством сравнения с показателями, полученными другими людьми по тому же тесту. При предметно-ориентированном тестировании, например, выполнение теста испытуемым описывается в единицах освоенных арифметических операций, объема словаря, уровня трудности доступного пониманию текста (от комиксов до литературной классики) или вероятности достижения определенного уровня выполнения деятельности в соответствии с внешним (профессиональным или образовательным) критерием.

До сих пор предметно-ориентированное тестирование находило применение главным образом в некоторых педагогических новшествах, и прежде всего, в программированном обучении, компьютеризованной профессиональной подготовке и других обучающих системах с выбором индивидуального темпа обучения. Во всех таких системах тестирование тесно интегрировано с обучением и проводится до, во время и после выполнения каждого учебного задания для проверки необходимых для обучения умений, выявления возможных трудностей усвоения материала и выбора последующих обучающих процедур (Nitko, 1989).

В другом ракурсе предметно-ориентированные тесты использовались в широких инспекторских проверках качества образования, таких как Национальная программа оценки прогресса в образовании (*National Assessment of Educational Progress*) (E. G. Johnson, 1992; Messick, Beaton, & Lord, 1983; F. B. Womer, 1970). Кроме того, они оказались полезными в удовлетворении запросов учебной отчетности. Еще одной иллюстрацией применения предметно-ориентированного тестирования могут служить экзамены на право вождения автомобиля или управление самолетом. Родственной областью является проверка профессиональной квалификации, где оценивается владение небольшим числом строго определенных профессиональных навыков, как это имеет место во многих военных специальностях (Maier, & Hirshfeld, 1978; Swezey, & Pearlstein, 1975).

Наконец, отметим, что знакомство с принципами предметно-ориентированного тестирования может способствовать усовершенствованию традиционных, неформальных тестов, составляемых учителями для использования в своем классе. Линн и Гронлунд (Linn, & Gronlund, 1995) разработали полезное руководство по составлению и проведению таких тестов с простым и хорошо построенным введением в предметно-ориентированное тестирование. Краткое, но превосходное обсуждение главных ограничений предметно-ориентированных тестов дано в работе Ибела (Ebel, 1972). Всестороннее рассмотрение многих специальных вопросов конструирования и оценивания таких тестов можно найти в руководстве под редакцией Берка (Berk, 1984a).

Значение содержания. Главным отличительным признаком предметно-ориентированного тестирования (как бы оно ни определялось и под каким бы названием ни выступало) является интерпретация выполнения теста с точки зрения его смыслового содержания. Упор делается на то, *что* тестируемые могут делать и *что* они знают, а не на то, как они выглядят на фоне других. Главное требование, которое необходимо соблюдать при конструировании теста этого типа, состоит в четком определении об-

ласти знаний или умений, которые предполагается оценивать с его помощью. Если мы хотим, чтобы показатели такого теста обладали поддающимся передаче значением, необходимо выбрать содержательную область, признаваемую всеми как важную. Выбранная область подразделяется затем на небольшие единицы, определяемые в терминах выполняемой деятельности. В контексте школьного обучения такие единицы соответствуют определяемым на поведенческом уровне учебным целям типа «умножить трехзначное число на двузначное» или «указать слово, в котором ошибочно написана буква *е* в суффиксе *ян*». В программах индивидуализированного обучения число таких целей-задач может достигать нескольких сотен по одному только учебному предмету. После того как все учебные цели сформулированы, нужно составить конкретные задания, обеспечивающие выборочную проверку достижения каждой из этих целей. По общему признанию, процедура эта достаточно трудна и поглощает много времени. Но без тщательной спецификации и контроля содержания заданий результаты предметно-ориентированного тестирования легко могут превратиться в чуждую и не поддающуюся интерпретации мешанину цифр. Возможный компромисс состоит в том, чтобы выявить и определить основные понятия, принципы, методы или учебные цели, прибегнув к помощи экспертов; затем каждую из определенных таким образом значимых областей можно тщательно проверить с помощью набора подходящих тестовых заданий. Безусловно, степень специфичности, с которой должны оцениваться области поведения, варьирует в зависимости от характера и цели теста (Popham, 1984; Roid, 1984).

Предметно-ориентированное тестирование, при правильном применении, лучше всего приспособлено для проверки базовых умений и навыков (таких, как навыки чтения и оперирования числами) на элементарных уровнях. В этих областях учебные цели-задачи обычно можно упорядочить в иерархическую последовательность, когда приобретение более элементарных навыков является предпосылкой для формирования навыков более высокого уровня.[1] Однако применительно к более высоким уровням знаний в сравнительно мало структурированных областях практически невозможно, да и нежелательно, формулировать такие цели с предельной конкретностью. На этих уровнях как само содержание, так и последовательность его усвоения, вероятно, должны определяться более гибко.

С другой стороны, делая акцент на содержании при интерпретации тестовых показателей, предметно-ориентированное тестирование может оказать благотворное влияние на тестирование в целом. От такого подхода выиграла бы, например, интерпретация показателей тестов интеллекта. Если выполнение ребенком теста интеллекта описывать исходя из специфических интеллектуальных умений и знаний, предполагаемых набором тестовых заданий, то это могло бы помочь в преодолении тех недоразумений и неправильных представлений, которыми к настоящему времени оброс традиционный *IQ*. Однако, когда предметно-ориентированный подход формулируется в этих общих выражениях, он равносилен интерпретированию тестовых показателей в свете подтвержденной валидности конкретного теста, а не в единицах каких-то туманных внутренних сущностей. Разумеется, такая интерпретация может комбинироваться с показателями, ориентированными на статистические нормы.

---

[1] В идеале такие тесты описываются симплексной моделью шкалы Гуттмана (см. Popham, & Husek, 1969), так же как и порядковые шкалы Пиаже, обсуждаемые в главе 9.

Тестирование овладения знаниями, умениями и навыками. Вторым важным признаком, обычно связываемым с предметно-ориентированным тестированием, является способ проверки овладения предметом. По существу, этот способ дает оценку по принципу «все или ничего», показывая, достиг или не достиг испытуемый заранее установленного уровня владения определенным предметом. При тестировании базовых умений и навыков этот уровень предполагает почти совершенное владение (требуя, например, правильного выполения 80–85 % всех заданий). Возможно также применение трехступенчатой шкалы, фиксирующей совершенное владение, невладение и промежуточный («критический») интервал, или интервал неопределенности.

В связи с индивидуализацией обучения некоторые педагоги пришли к убеждению, что при условии достаточного количества времени и адекватных методов обучения почти каждый может полностью справиться с поставленными перед ним учебными целями-задачами. В этом случае индивидуальные различия будут проявляться скорее во времени научения, чем в конечном результате, как при традиционном образовательном тестировании (Carroll, 1963, 1970; Cooley, & Glaser, 1969; Gagné, 1965). Из этого следует, что при тестировании овладения предметом индивидуальные различия в выполнении теста не представляют никакого или почти никакого интереса. В результате, предметно-ориентированные тесты в том виде, как они обычно конструируются, минимизируют индивидуальные различия в выполнении теста после соответствующего обучения. Тестирование овладения предметом систематически используется в упоминавшихся выше программах индивидуализированного обучения. На этих же принципах построены регулярно издаваемые предметно-ориентированные тесты базовых умений и навыков, пригодные для младших и средних классов школы.

При конструировании таких тестов встают два важных вопроса: 1) Сколько заданий нужно включить в тест для надежной оценки достижения каждой из конкретных учебных целей? 2) Какая доля заданий должна быть выполнена правильно для надежного установления владения предметом? На начальных этапах развития предметно-ориентированного тестирования ответы на эти вопросы опирались на субъективное мнение. Со временем, однако, был достигнут существенный прогресс в разработке статистических методов, позволяющих давать на них объективные, эмпирически обоснованные ответы (Berk, 1984a; R. L. Ferguson, & Novick, 1973; Hambleton, 1884a, 1989; Hambleton, & Novick, 1973). Несколько примеров помогут наглядно представить характер и диапазон этих разработок.

Эти два вопроса — о количестве заданий и граничных значениях показателя — можно объединить в одну гипотезу, поддающуюся проверке в рамках теории принятия решения и последовательного анализа (Hambleton, 1984a; Wald, 1947). Конкретно, мы хотим проверить гипотезу о том, что тестируемый достиг конкретной учебной цели или, иначе говоря, требуемого уровня владения определенным предметом, представленным набором заданий теста. Последовательный анализ состоит в проведении наблюдений, по одному за раз, и решении после каждого из них, следует ли 1) принять гипотезу, 2) отклонить гипотезу или 3) продолжать наблюдения. Таким образом, число наблюдений (в данном случае, число заданий), необходимых для получения надежного вывода, само определяется в процессе тестирования. Вместо того чтобы работать с фиксированным, заранее установленным числом заданий, экзаменуемый продолжает выполнять тест до тех пор, пока не будет принято решение о владении или невладении предметом. В этот момент тестирование прекращается, и учащийся либо переводится на следующий уровень обучения, либо возвращается к неосвоенному

уровню для дополнительного изучения. С учетом описанных выше в этой главе возможностей компьютеров, такие последовательные процедуры принятия решений стали практически осуществимыми и могут сокращать суммарное время тестирования, обеспечивая надежные оценки овладения той или иной предметной областью.

Некоторые исследователи изучают возможности оценивания владения предметом на основе байесовских методов, позволяющих учитывать косвенные данные и идеально подходящих для принятия решений такого рода, которые требуются при тестировании уровня овладения знаниями, умениями и навыками. Из-за большого количества конкретных учебных целей, достижение которых должно оцениваться, в предметно-ориентированных тестах на каждую такую цель обычно приходиться лишь небольшое число заданий. Для дополнения этой ограниченной информации и были разработаны методы, учитывающие косвенные данные о прежних достижениях ученика, а также о результатах тестирования других учащихся (R. L. Ferguson, & Novick, 1973; Hambleton, 1984a; Hambleton, & Novick, 1973).

Когда невозможно применение индивидуально адаптируемых методик, граничные значения показателей могут устанавливаться эмпирически, на основе анализа показателей по данному тесту, получаемых подходящими группами до и после обучения. В этом случае граничное значение выбирается таким образом, чтобы наилучшим образом дифференцировать получивших и не получивших соответствующее обучение (Panell, & Laabs, 1979; L. A. Shepard, 1984). В специфических ситуациях требуется дополнительный анализ на предмет оценки относительной серьезности «прохождения» теста теми, кто не обучался, и, напротив, «непрохождения» теста теми, кто получил требуемое обучение. Граничное значение показателя можно было бы соответственно повысить или понизить, чтобы привести в соответствие с последствиями ошибочной классификации.

Связь с тестированием, ориентированным на нормы. За пределами базовых умений и навыков тестирование владения предметом неприменимо или недостаточно. В более сложных и менее структурированных областях не существует предела достижений. Конкретный человек может почти неограниченно совершенствовать такие функции, как понимание, критическое мышление, предчувствие и оригинальность. Кроме того, усвоение содержания может идти различными путями в зависимости от способностей, интересов и целей человека, а также от местных образовательных возможностей. При этих условиях совершенное владение нереально, да и не нужно. Вот почему в таких случаях обычно применяется ориентированное на нормы оценивание степени образованности или квалификации. Некоторые издаваемые тесты построены таким образом, что допускают как предметно-ориентированное, так и ориентированное на статистические нормы применение. Примером могут служить стэнфордские диагностические тесты чтения и математики. Обеспечивая соответствующие нормы на каждом уровне, эти тесты позволяют проводить качественный анализ достижения ребенком детализированных учебных целей.

Следует заметить, что предметно-ориентированное тестирование вовсе не так ново и не столь уж сильно отличается от ориентированного на статистические нормы тестирования, как полагают некоторые из его сторонников. Оценка индивидуального выполнения теста в абсолютных единицах, таких как буквенные отметки (*letter grades*) или процент правильных ответов, несомненно, намного старше нормативной интерпретации. Еще до введения термина «критериально-ориентированное тестирова-

ние» делались попытки более точно описать выполнение теста с точки зрения его содержания (Ebel, 1962; J. C. Flanagan, 1962; Nitko, 1984, p. 14–16). Другие примеры можно найти среди первых шкал для оценивания качества почерка, сочинений или рисунков на основе сопоставления образцов работы индивидуума с набором стандартных образцов. Более того, как заметил Ибел (Ebel, 1972), в педагогике понятие овладения (*mastery*) чем-либо — в смысле усвоения определенных учебных единиц по принципу «все или ничего» — достигло значительной популярности в 1920–1930-х гг., но позднее от него отказались.

Нормативная основа имплицитно присутствует во всяком тестировании, независимо от того, как выражаются показатели теста (Angoff, 1974; Nitko, 1984). Сам выбор содержания или навыков, подлежащих измерению, определяется знанием специалиста, чего можно ожидать от людей на определенном уровне их развития или обучения. Такой выбор предполагает наличие сведений о том, как другие действовали в подобных ситуациях. Кроме того, устанавливая единые граничные значения показателя на континууме умения, тестирование овладения предметом не устраняет индивидуальных различий. Например, если уровень понимания текста задается формулировкой «умение понять содержание газеты “Нью-Йорк Таймс”», то все еще остается достаточно места для значительных индивидуальных различий в степени понимания. Применяя критический балл для дихотомизации выполнения теста, мы просто игнорируем индивидуальные различия, сохраняющиеся в рамках двух устанавливаемых категорий, и тем самым отбрасываем потенциально полезную информацию.

Минимальные квалификационные требования и критические показатели

Практические потребности и подводные камни. Понятие овладения (*mastery*) в предметно-ориентированном тестировании — это только один пример использования критических показателей в принятии решения. Повседневная жизнь обязывает точно формулировать и выполнять минимальные квалификационные требования к человеческой деятельности в самых различных областях. Во многих ситуациях соображения безопасности требуют установления критических, граничных точек в исполнении деятельности, как при выдаче водительских прав, отборе летчиков гражданской авиации или найме рабочих для обслуживания ядерных установок. В области образования прохождение университетского курса или окончание школы представляют собой другие ситуации, которые также требуют классификации людей по принципу «все или ничего» (Jaeger, 1989). В клинической и консультационной практике решения, касающиеся выбора лечения или линии поведения, могут требовать аналогичных, дихотомических, оценок.

Особо сильный довод в пользу применения граничных показателей связан с наличием критических переменных, необходимых для выполнения некоторых функций. Критическими называют такие переменные, недостаток в которых не может быть компенсирован выдающимися способностями или высочайшей квалификацией в областях, связанных с другими параметрами деятельности. В таких случаях высокий показатель по комплексной батарее профотбора мог бы маскировать недостаток критического умения. Однако при использовании граничных значений все те, кто не набрал требуемого минимума баллов по критическому умению, считаются не прошедшими

отбор, независимо от их других способностей и умений. Например, гидроакустики должны обладать высокой слуховой различительной чувствительностью. Во время Второй мировой войны новобранцев ВМФ США первоначально отбирали для обучения специальности гидроакустика на основе их совокупных показателей по тестам слухового различения и понимания механических закономерностей. В результате, целый ряд мужчин, обучавшихся до войны в колледже и потому сведущих в механике, но, к сожалению, не обладавших требуемым уровнем развития слухового различения, был зачислен на курсы гидроакустиков, с последующим отсевом. В соответствии с заведенным в ВМФ порядком несправившихся с первым учебным заданием переводили на неквалифицированную работу — учениками матросов, теряя в связи с этим возможность использовать их в качестве специалистов. Дополнительный анализ сложившегося положения привел со временем к замене критерия отсева в процедуре отбора по этой военной специальности. Однако для большинства имеющих отношение к работе переменных их связь с эффективностью труда носит линейный характер, так что чем выше показатель по тесту, тем лучше, в общем, человек справляется с работой (Coward, & Sackett, 1990). В таких случаях, фактический показатель человека по соответствующему тесту является лучшим прогнозирующим параметром, чем его положение относительно граничной точки.

Коль скоро невозможно избежать использования критических показателей при принятии многих практических решений, важно сознавать подводные камни таких оценок и применять меры для сокращения ошибочных решений. Например, нужно стремиться смягчать ограничивающее действие единственного тестового показателя. Когда это возможно, следует предпочесть критический интервал или группу критических показателей одному-единственному показателю, полученному при однократном проведении конкретного теста. Кроме того, решения, принимаемые в отношении конкретных лиц, должны основываться на информации из разных источников, дополняющих тестовые показатели другими релевантными данными в отношении интересующей деятельности в прошлом и настоящем. Если граничные значения показателей по тестам устанавливаются группой экспертов, в ней должно быть обеспечено адекватное представительство специалистов как в области предполагаемой профессиональной деятельности, так и в области конструирования и применения тестов. Самое главное, при появлении возможности граничные значения показателей следует определять или верифицировать на основе эмпирических данных. В частности, это предполагает получение тестовых показателей на группах, которые явно различаются по критерию релевантного поведения, такому как фактическое выполнение данного вида работы. Разумеется, именно это выполнение и предназначен предсказывать конкретный тест, критический показатель по которому должен гарантировать безопасный, приемлемый или желаемый минимум. Ясной иллюстрацией эмпирического метода установления критических показателей по тесту для отбора персонала служат таблицы ожидаемых результатов (*expectancy tables*), рассматриваемые в следующем разделе.

Таблицы ожидаемых результатов. Результаты теста можно также интерпретировать опираясь на критерий ожидаемого выполнения предстоящей программы обучения или работы. Такое употребление термина «критерий» соответствует сложившейся в психометрии традиции, как в тех случаях, когда говорят, что валидность теста устанавливается относительно некоторого критерия (см. главу 1). Строго говоря, термин «критериально-ориентированное тестирование» следовало бы использовать при-

менительно к этому типу интерпретации выполнения теста, тогда как другие подходы, обсуждавшиеся в предыдущем разделе, правильнее было бы характеризовать как содержательно- или предметно-ориентированные.

В таблице ожидаемых результатов приводятся вероятности различных критериальных исходов для лиц, получивших тот или иной тестовый балл. Например, если учащийся набрал 530 баллов по Тесту академической оценки (*SAT*) Совета колледжей, то каковы его шансы закончить первый курс определенного колледжа со средней оценкой *A*, *B*, *C*, *D* или *F*? Информацию такого рода можно получить, изучая двумерное распределение, связывающее значения прогнозирующих показателей (*SAT*) с критерием статуса студента первого курса (средней оценкой успеваемости). Если число случаев в каждой ячейке такого двумерного распределения заменить на проценты, получится таблица ожидаемых результатов, такая как табл. 3–6. В ней представлены данные, полученные при обследовании 211 учащихся 7-х классов, записавшихся на курс математики. В качестве предиктора здесь использован тест числового рассуждения из Дифференциальных тестов способностей (*DAT*), проведенный в конце первого семестра, а в качестве критерия — итоговые оценки по курсу математики в конце второго семестра. Корреляция между тестовыми показателями и критерием составила 0,60.

Таблица 3–6

Таблица ожидаемых результатов, демонстрирующая связь между показателями теста числового рассуждения (из *DAT*) и итоговыми оценками по курсу математики 211 учащихся 7-х классов

Тестовый показатель	Число случаев	Процент получивших каждую оценку			
		D и ниже	*C*	*B*	*A*
30 и выше	22	5	0	36	59
20–29	104	9	21	43	27
10–19	71	36	37	24	3
Ниже 10	14	43	36	·14	7

(С упрощениями из Technical Manual for Differential Aptitude Tests, *5th ed., p. 152. Воспроизведено с разрешения Психологической корпорации. Copyright © 1992 by The Psychological Corporation)*

В первой колонке табл. 3–6 приведены тестовые показатели, сгруппированные в четыре интервала, во второй — число учащихся, тестовые показатели которых попали в соответствующий интервал. Остальные цифры таблицы (по строкам) показывают процент учащихся внутри каждого интервала группирования показателей теста, получивших оценку *A*, *B*, *C* или *D* (и ниже) по окончании курса. Так, из 22 учеников, набравших в тесте числового рассуждения 30 и более баллов, 5 % получили оценку *D* (или ниже), никто не получил оценку *C*, 36 % получили оценку *B* и 59 % — оценку *A*. На другом краю распределения, из 14 учеников с тестовым показателем ниже 10 баллов получили оценку *D* (или ниже) 43 %, *C* — 36 % и *B* — 14 %. Аномальные 7 % учеников, получивших оценку *A*, представляют собой лишь один случай и потому не несут практически полезной информации для обобщения, так же как и 5 % учеников с тестовым показателем 30 (и более) баллов, получивших оценку *D* (или ниже), опять-таки представленных одним случаем. Тем не менее с учетом ограничений имеющихся данных, проценты в табл. 3–6 дают оценки вероятности получения индивидуумом

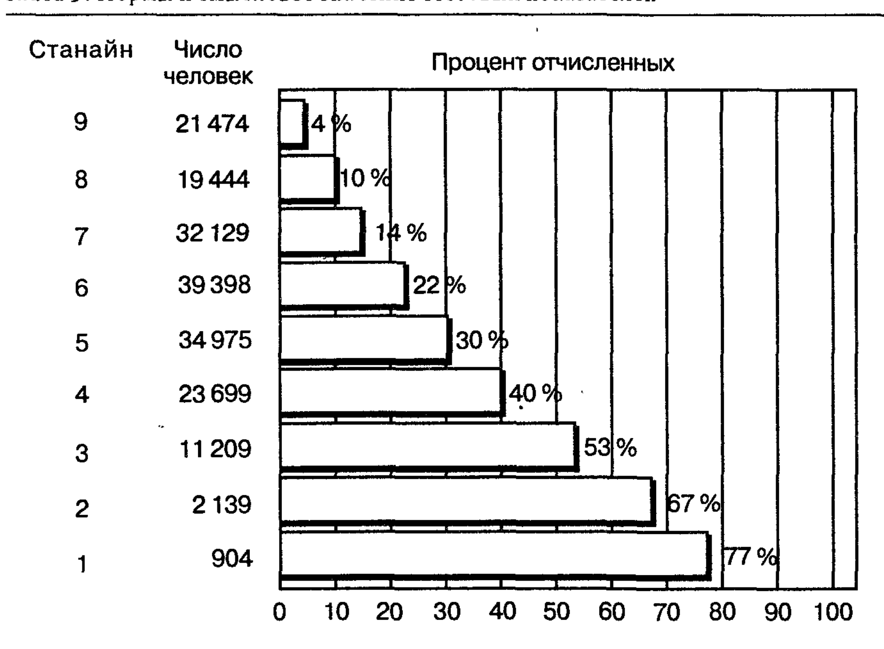


Рис. 3–7. Диаграмма ожидаемого отсева, показывающая связь между выполнением заданий батареи отбора летчиков и отчислением с начального курса летной подготовки

(Из J. C. Flanagan, 1947, p. 58)

того или иного критериального балла. Например, если новый ученик наберет 24 балла по тесту числового рассуждения *DAT* (т. е. попадет в интервал группирования 20–29), его шансы получить *A* по курсу математики можно оценить как 27 из 100, а шансы получить *B* — как 43 из 100, и т. д.

Во многих практических ситуациях может отдаваться предпочтение дихотомическим критериям в виде «успеха» или «неудачи» в работе, в прохождении учебного курса и т. д. В этих условиях можно построить *диаграмму ожидаемого отсева*, показывающую вероятность успеха или неудачи для каждого интервала группирования тестовых показателей. Рис. 3–7 дает пример такой диаграммы. Базирующаяся на батарее отбора летчиков, разработанной ВВС США, эта диаграмма ожидаемого отсева показывает для каждого станайна шкалы процент курсантов, не справившихся с начальным курсом летной подготовки. Можно увидеть, что в процессе подготовки отсеялись 77 % курсантов, получивших тестовый показатель, равный 1 станайну, и только 4 % курсантов, получивших показатель, равный 9 станайнам. Между этими крайними значениями процент неудач неуклонно снижается с прибавлением каждого станайна. На основе этой диаграммы ожидаемого отсева можно было бы предсказать, например, что приблизительно 40 % курсантов с тестовым показателем, равным 4 станайнам, потерпят неудачу и приблизительно 60 % из них удовлетворительно завершат начальный курс летной подготовки. Аналогичные прогнозы по каждому станайну можно строить и относительно вероятности успеха или неудачи отдельных курсантов. Так, получив тестовый показатель, равный 4 станайнам, курсант имеет 60 шансов против 40, т. е. 3 шанса против 2, успешно закончить начальный курс летной подготовки. Нетрудно видеть, что помимо обеспечения критериально-ориентированной интерпретации тес-

товых показателей таблицы ожидаемых результатов и диаграммы ожидаемого отсева дают общее представление о валидности теста в предсказании по данному критерию. По этой причине эмпирические процедуры установления критических значений тестового показателя более подробно обсуждаются в конце главе 6, в разделе о моделях принятия решений в честном использовании тестов. В этом разделе также упоминаются математические методы для установления оптимальных критических значений тестового показателя при различных условиях. Кроме того, с конкретными приложениями критических показателей в основных областях психологической практики можно ознакомиться в главе 17.

4 НАДЕЖНОСТЬ

Под надежностью понимается устойчивость, или согласованность (*consistency*) результатов теста, получаемых при повторном его применении к тем же испытуемым в различные моменты времени, при использовании разных наборов эквивалентных заданий или же при изменении других условий обследования. Такое понимание надежности лежит в основе вычисления *ошибки измерения* отдельного показателя, благодаря чему мы можем предсказывать диапазон случайных колебаний тестового балла у конкретного человека, возникающих, вероятно, под действием посторонних или неизвестных факторов.

Понятие надежности обычно охватывает несколько аспектов устойчивости тестовых показателей. В самом широком смысле надежность теста показывает, в какой степени индивидуальные различия в тестовых показателях могут быть отнесены на счет «истинных» различий в изучаемых свойствах, а в какой могут быть приписаны случайным ошибкам. Говоря более специальным языком, меры надежности теста позволяют оценить, какую долю общей дисперсии (общей изменчивости) тестовых показателей составляет *дисперсия ошибок*. Это не «ошибки» в обычном смысле слова, предполагающем, что их можно было бы избежать или скорректировать путем усовершенствования методологии измерений. Данное терминологическое значение слова «ошибка» унаследовано из более ранней эпохи в развитии психологии, когда интерес ученых сосредоточивался на выявлении общих законов поведения и оценивании испытуемых по таким свойствам, которые считались неизменными базовыми чертами. В наше время психологи признают изменчивость существенным свойством всякого поведения и потому занимаются выявлением и классификацией многочисленных источников такой изменчивости.

Что касается надежности показателя, суть дела заключается в определении дисперсии ошибок. Факторы, которые применительно к одним задачам можно было бы счесть источниками случайной вариации показателя (т. е. дисперсии ошибок), при решении других задач могут быть отнесены, и не без основания, к причинам его истинной дисперсии. Например, если бы нас интересовало измерение колебаний настроения, то происходящие день ото дня изменения в показателях шкалы «радость — уныние» были бы релевантны цели данного теста и, следовательно, составляли бы часть

истинной дисперсии показателей. С другой стороны, если бы тест предназначался для измерения более устойчивых характеристик личности, те же ежедневные колебания попали бы уже в разряд дисперсии ошибок.

В сущности, любое условие тестирования, которое не имеет отношения к цели теста, представляет собой источник дисперсии ошибок. Поэтому, стремясь к поддержанию единых условий тестирования (контролируя общую обстановку, временные ограничения, инструкции испытуемым, раппорт и другие аналогичные факторы), пользователи тестов способствуют уменьшению дисперсии ошибок и повышению надежности тестовых показателей. Но и при оптимальных условиях тестирования ни один тест не является абсолютно надежным инструментом. Поэтому каждый тест следует сопровождать сведениями о его надежности. Сообщаемая мера надежности характеризует тест только в случае его проведения в стандартных условиях и с людьми, имеющими сходство с теми, кто входил в состав нормативной выборки. Следовательно, при описании теста нужно точно указывать и характеристики этой выборки, вместе с типом измеренной на ней надежности.

Теоретически, разновидностей тестовой надежности может быть очень много — столько же, сколько и условий, влияющих на показатели теста, так как любое из этих условий может оказаться нерелевантным конкретной цели тестирования и потому отнесенным к источникам дисперсии ошибок. Однако практическое применение находит лишь несколько типов надежности. В этой главе мы обсудим основные способы измерения надежности тестовых показателей, вместе с источниками дисперсии ошибок, идентифицируемыми каждым из этих способов.[1]

Поскольку все типы надежности касаются степени согласованности или соответствия между двумя независимо полученными множествами показателей, их все можно выразить в виде *коэффициента корреляции.* Соответственно, с целью разъяснить использование и интерпретацию коэффициентов корреляции, в следующем разделе рассматриваются их основные характеристики. Более специальное обсуждение корреляции, с подробным описанием вычислительных процедур, можно найти в любом элементарном учебнике по статистике для педагогов и психологов (см, например, Runyon, & Haber, 1991; D. C. Howell, 1997).

Коэффициент корреляции

Смысл корреляции. По существу, коэффициент корреляции (*r*) выражает степень соответствия или *связи* между двумя множествами показателей. Например, если испытуемый, получивший высший показатель по переменной 1, получает высший показатель и по переменной 2, а испытуемый, получивший второй лучший показатель по переменной 1, получает такой же показатель по переменной 2 и т. д. до самого низшего

---

[1] Этот подход к надежности показателей иногда называли теорией надежности как обобщаемости (см. Brennan, 1994; Crick & Brennan, 1982; Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Feldt, & Brennan, 1989; Shavelson & Webb, 1991). Однако это название недостаточно специфично для дифференциального термина, так как понятие обобщаемости применимо ко всем аспектам тестовых показателей, да и, фактически, ко всем научным данным. Более точная характеристика этого метода определения надежности основана на его способности идентифицировать компоненты дисперсии как релевантные или нерелевантные.

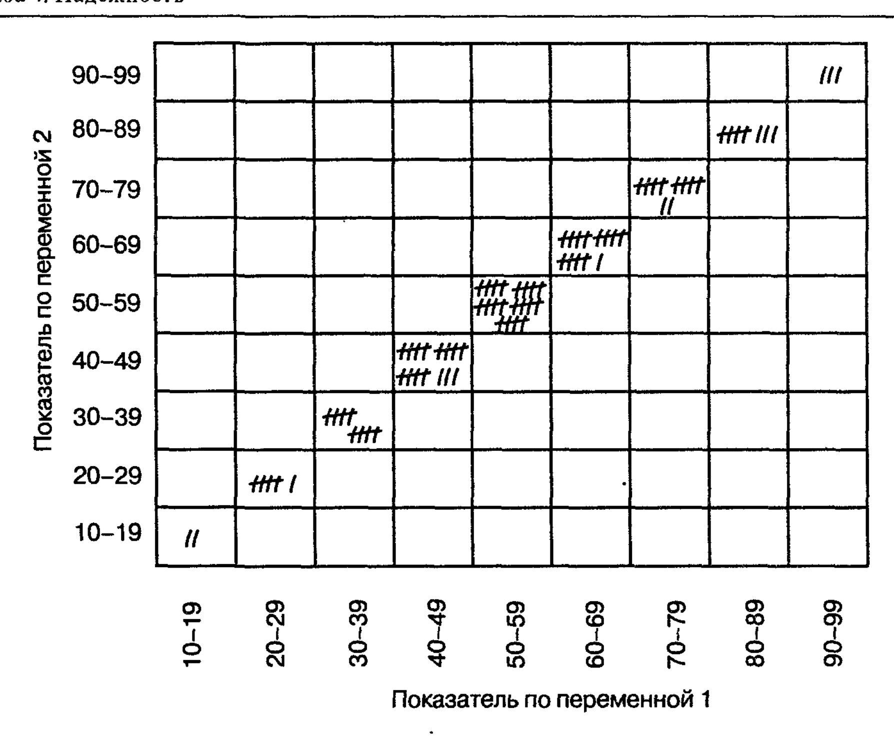


Рис. 4–1. Двумерное распределение для гипотетической корреляции (+ 1)

показателя в группе, то имеет место прямолинейная корреляция между переменными 1 и 2. Величина корреляции составляет в этом случае + 1,0.

Рис. 4–1 иллюстрирует гипотетический случай прямолинейной положительной корреляции. На рисунке представлена диаграмма рассеяния, или двумерное распределение. Каждая палочка на этой диаграмме отмечает показатель испытуемого как по переменной 1 (горизонтальная ось), так и по переменной 2 (вертикальная ось). Нетрудно заметить, что все 100 случаев в данной группе распределились вдоль диагонали, идущей из левого нижнего угла в правый верхний угол диаграммы. Такое распределение означает прямолинейную положительную корреляцию (+ 1,00), поскольку из него видно, что относительное положение каждого испытуемого по обеим переменным одинаково. На практике, чем ближе двумерное распределение показателей к этой диагонали, тем выше положительная корреляция между ними.

На рис. 4–2 изображена прямолинейная отрицательная корреляция (– 1,00). В этом случае имеет место полная инверсия показателей по двум переменным: лучший индивидуальный результат по переменной 1 соответствует худшему по переменной 2, и наоборот, причем это обратное соотношение показателей сохраняется неизменным на всем распределении. Из диаграммы рассеяния видно, что все испытуемые распределяются по диагонали, идущей из левого верхнего в правый нижний угол.

Нулевая корреляция указывает на полное отсутствие связи. Если положение каждого испытуемого относительно переменной 1 определить методом вытаскивания бумажек с именами из шляпы, а затем ту же процедуру повторить для переменной 2, то в итоге мы и получим нулевую или близкую к нулю корреляцию. При этих условиях, зная результат индивидуума по переменной 1, невозможно предсказать его относи-

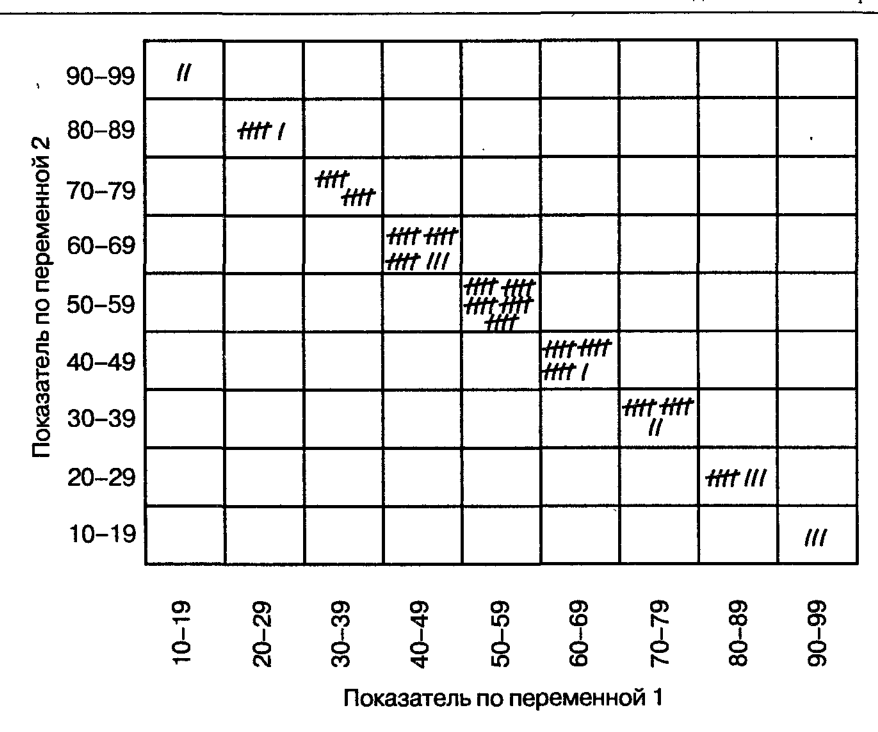


Рис. 4—2. Двумерное распределение для гипотетической корреляции (– 1)

тельное положение на переменной 2. Испытуемый, имеющий высший показатель по переменной 1, мог бы получить высокий, средний или низкий показатель по переменной 2. Одни испытуемые могут случайно оказаться выше или ниже среднего показателя по обеим переменным, другие будут выше среднего по одной переменной и ниже среднего по другой, иными словами, не будет никакой закономерности в связи показателей при переходе от одного человека к другому.

Вычисляемые по реальным данным коэффициенты корреляции попадают между граничными значениями (– 1 и + 1) и обычно отличаются от нуля, но практически всегда оказываются меньше единицы (по абсолютному значению). Корреляция между показателями способностей почти всегда положительна, хотя часто невысока. Когда между двумя такими переменными обнаруживается отрицательная корреляция, обычно это результат того, каким способом выражались показатели по этим переменным. Например, если временны́е показатели коррелировать с показателями суммарной результативности, то результатом, скорее всего, будет отрицательная корреляция. Так, если показатель каждого испытуемого по тесту арифметических вычислений выражается количеством минут, затраченных на выполнение всех заданий, тогда как показатель по тесту арифметических рассуждений представлен числом правильно решенных задач, то можно ожидать появления отрицательной корреляции между этими показателями. В данном случае наименее успевающий (работающий медленнее всех) испытуемый получит численно самый высокий показатель по первому тесту, в то время как по второму тесту самый высокий показатель будет у наиболее успевающего, т. е. решившего больше всего задач, испытуемого.

Коэффициенты корреляции можно вычислять разными способами, в зависимости от природы данных. Наибольшее распространение получил *коэффициент корреляции произведения моментов Пирсона.* Этот коэффициент учитывает не только положение индивидуума в группе, но и степень его отклонения в ту или иную сторону от среднего уровня группы. Напомним, что когда положение каждого индивидуума выражается в единицах стандартных показателей, те, кто занимает положение выше среднего, получают положительные стандартные показатели, а те, кто находится ниже среднего уровня, — отрицательные. Таким образом, испытуемый, превосходящий группу по уровню обеих коррелируемых переменных, будет иметь два положительных стандартных показателя, а испытуемый, отстающий от группы по уровню этих переменных, — два отрицательных. Если теперь перемножить стандартные показатели каждого из этих испытуемых по обеим переменным, то оба произведения будут положительны. Пирсоновский коэффициент корреляции есть просто среднее арифметическое всех таких произведений. Его числовое значение бывает высоким и положительным, когда соответствующие стандартные показатели имеют по обеим переменным одинаковые знаки и приблизительно равную величину. Когда испытуемых занимают положение выше среднего по одной переменной, но ниже среднего по другой, то соответствующие произведения будут отрицательны. А если сумма произведений отрицательна, то отрицательной будет и корреляция. Когда же одни произведения отрицательны, а другие положительны, корреляция будет близка к нулю.

На практике нет необходимости переводить каждый первичный показатель в стандартный перед нахождением их произведений, так как это преобразование можно выполнить разом для всех показателей после суммирования их попарных произведений. Существует много ускоренных методов вычисления коэффициента корреляции Пирсона. Метод, представленный в табл. 4–1, не самый быстрый, но зато он лучше других раскрывает смысл коэффициента корреляции. В табл. 4–1 показано вычисление r Пирсона между показателями по арифметическому тесту и тесту чтения у 10 детей. В двух столбцах справа от имен учеников приведены их показатели по первому (X) и второму (Y) тесту. Суммы и средние арифметические 10 показателей приведены под соответствующими столбцами. В третьем столбце приведены отклонения (x) каждого показателя по арифметическому тесту от среднего арифметического этих показателей, а в четвертом — отклонения (y) индивидуальных показателей по тесту чтения от их среднего арифметического. Квадраты этих отклонений даны в следующих двух столбцах таблицы, а суммы квадратов отклонений используются при вычислении стандартных отклонений показателей по обоим тестам с помощью метода, описанного в главе 3. Вместо того чтобы каждое x и y делить на соответствующее SD для получения стандартных показателей, это деление выполняется только раз, в конце, как показано в формуле коэффициента корреляции в нижней части табл. 4–1. Попарные произведения (xy) в последнем столбце получены перемножением соответствующих отклонений в столбцах (x) и (y). Для вычисления корреляции (r) сумма этих попарных произведений делится на число случаев (N) и на произведение двух стандартных отклонений ($SD_x SD_y$).

Статистическая значимость. Вычисленная в табл. 4–1 корреляция ($r = 0{,}40$) указывает на умеренную положительную связь между показателями арифметического теста и теста чтения. То есть налицо некоторая тенденция, выражающаяся в том, что дети, хорошо показавшие себя в арифметическом тесте, также неплохо справляются с тес-

Таблица 4–1

Вычисление коэффициента корреляции произведения моментов Пирсона

Ученик	Арифметика X	Чтение Y	x	y	x^2	y^2	xy
Билл	41	17	+ 1	– 4	1	16	– 4
Кэрол	38	28	– 2	+ 7	4	49	– 14
Джеффри	48	22	+ 8	+ 1	64	1	8
Энн	32	16	– 8	– 5	64	25	40
Боб	34	18	– 6	– 3	36	9	18
Джейн	36	15	– 4	– 6	16	36	24
Элен	41	24	+ 1	+ 3	1	9	3
Рут	43	20	+ 3	– 1	9	1	– 3
Дик	47	23	+ 7	+ 2	49	4	14
Мери	40	27	0	+ 6	0	36	0
Σ	400	210	0	0	244	186	86
M	40	21					

$$SD_x = \sqrt{\frac{244}{10}} = \sqrt{24{,}40} = 4{,}94 \qquad SD_y = \sqrt{\frac{186}{10}} = \sqrt{18{,}60} = 4{,}31$$

$$r_{xy} = \frac{\sum xy}{N \cdot SD_x \cdot SD_y} = \frac{86}{10 \cdot 4{,}94 \cdot 4{,}31} = \frac{86}{212{,}91} = 0{,}40$$

том чтения, и наоборот. Если нас интересуют результаты только этих 10 детей, мы можем принять полученный коэффициент корреляции в качестве адекватной характеристики степени связи, существующей между двумя переменными в данной группе. В психологических исследованиях, однако, обычно стремятся распространить полученный на частной *выборке* испытуемых результат на более широкую *совокупность*, представленную этими испытуемыми. Например, мы могли бы задаться вопросом, существует ли связь между арифметическими навыками и навыками чтения у американских школьников того же возраста, что и наши испытуемые. Конечно, 10 исследованных случаев — совершенно недостаточная выборка из такой совокупности, ибо на другой сравнимой по размерам выборке можно было бы получить как гораздо более низкую, так и значительно более высокую корреляцию.

Существуют статистические методы оценки вероятных колебаний от одной выборки к другой коэффициентов корреляции, средних, стандартных отклонений и любых других групповых показателей. Вопрос, обычно задаваемый по поводу коэффициентов корреляции, еще проще: отличается ли выборочная корреляция существенно от нуля? Иными словами, если в генеральной совокупности корреляция равна нулю, то могла ли полученная на нашей выборке столь высокая корреляция появиться в результате одной только ошибки выборки? Когда говорят, что корреляция значима «на 1 %-ном уровне» (или «на уровне 0,01»), то имеют в виду следующее: существует всего лишь один шанс из ста, что в генеральной совокупности данный коэффициент равен нулю. Из чего можно сделать вывод, что между этими двумя переменными действительно имеет место корреляция. Уровни значимости указывают на приемлемую для исследователя степень риска совершить ошибку в выводах из полученных данных. Когда говорят, что корреляция значима на уровне 0,05, то вероятность ошиб-

ки составляет уже пять шансов из ста. В большинстве психологических исследований применяются 1 и 5 %-ный уровни значимости, хотя при необходимости или желании можно пользоваться и другими уровнями значимости.

Вычисленная в табл. 4–1 корреляция, равная 0,40, незначима даже на уровне 0,05, что вполне ожидаемо, поскольку по 10 случаям трудно вывести общую закономерность, касающуюся связи между переменными. Для выборки такого объема самая малая корреляция, значимая на уровне 0,05, равна 0,63. Любая корреляция ниже этой величины оставляет без ответа вопрос о том, коррелируют ли эти две переменные в совокупности, из которой была извлечена выборка. Минимальные значимые (на 1 и 5 %-ном уровнях) коэффициенты корреляции для выборок разного объема можно определить по справочным таблицам значимости коэффициентов корреляции, имеющимся в любом приличном учебнике статистики. Однако для понимания проблематики этой книги требуется лишь общее представление об основных статистических понятиях.

В течение многих лет уровни значимости были традиционным средством оценивания корреляций. Тем не менее сейчас все больше сознаются недостатки этого подхода и его несоответствие потребностям исследователей. Доказательство того, что коэффициент надежности (или любая корреляция) значимо отличается от нуля, мало что дает как для теории, так и для практики. Даже высокая корреляция, когда она получена на малой выборке, не удовлетворяет «критерию значимости». Приходящий на смену уровням значимости и завоевывающий все большее признание подход учитывает фактическую величину полученной корреляции и оценивает границы *доверительного интервала*, в который — на выбранном уровне доверительной вероятности — попадает значение генерального коэффициента корреляции (см., например, Carver, 1993; J. Cohen, 1994; Hunter, & Schmidt, 1990; Olkin, & Finn, 1995; Schmidt, 1996; W. W. Tryon, 1996). Это смещение интереса к доверительным интервалам как дополнению, если не замене проверки значимости, предвещает важный сдвиг в анализе коэффициентов корреляции в ближайшие годы.

Коэффициент надежности. Коэффициенты корреляции широко применяются в анализе психометрических данных. Одно из применений таких коэффициентов — это измерение надежности теста. Пример коэффициента надежности, вычисленного пирсоновским методом произведения моментов, приведен на рис. 4–3. В этом случае рассчитывалась корреляция между показателями 104 человек по двум эквивалентным формам теста «беглость речи».[1] В обоих случаях испытуемым давалось пять минут, в течение которых они должны были написать как можно больше слов, начинающихся на заданную букву. Формы теста отличались друг от друга лишь задаваемой буквой. Авторы теста подобрали начальные буквы с таким расчетом, чтобы трудность заданий была примерно одинаковой.

Корреляция между числом слов, написанных в ходе выполнения каждой из двух форм данного теста, оказалась равной 0,72, т. е. довольно высокой и значимой на уровне 0,01. При объеме выборки N = 104 любая корреляция от 0,25 и выше является значимой на этом уровне. И все же полученная корреляция несколько ниже, чем это желательно для коэффициентов надежности, обычно превышающих 0,80 и даже 0,90.

[1] Одного из субтестов Тестов первичных умственных способностей для возраста 11–17 лет, разработанных *SRA*. Данные получены в исследовании Анастази и Дрейка (Anastasi & Drake, 1954).

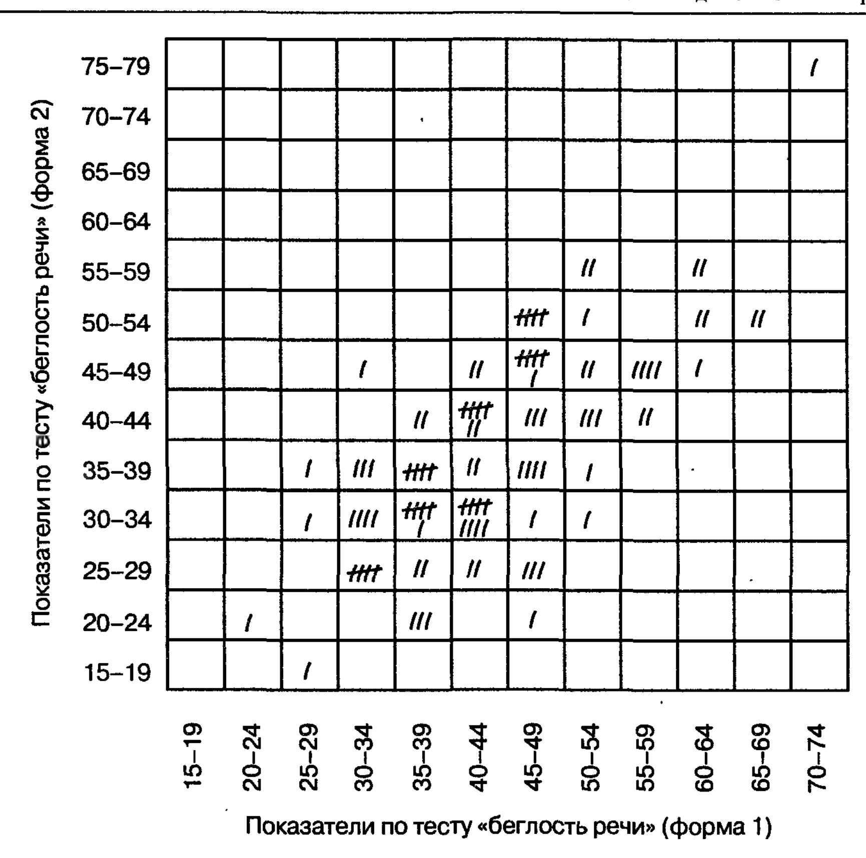


Рис. 4–3. Коэффициент надежности 0,72 *(по данным из статьи Anastasi & Drake, 1954).*

Диаграмма рассеяния для этих данных (рис. 4–3) представляет типичное двумерное распределение, соответствующее высокой положительной корреляции. Можно видеть, как «палочки» (условные значки для кодировки испытуемых или, в общем, наблюденных случаев) теснятся вблизи диагонали, идущей из левого нижнего в правый верхний угол; тенденция группировки в этом направлении выражена довольно определенно, хотя и наблюдается некоторый разброс отдельных случаев. В следующем разделе обсуждается использование коэффициента корреляции для вычисления различных мер надежности теста.

Типы надежности

Ретестовая надежность. Самый очевидный и понятный метод определения надежности результатов теста — его повторное проведение. В этом случае коэффициент надежности (r_n) просто равен корреляции между показателями, полученными теми же испытуемыми в каждом из двух случаев проведения теста. Дисперсия ошибок соответствует случайным колебаниям в выполнении заданий от одного сеанса тестирования к другому. Эти колебания могут отчасти быть результатом неконтролируе-

мых условий тестирования — таких, как резкие изменения погоды, внезапные шумы и другие отвлекающие факторы или, скажем, сломавшийся некстати карандаш. В какой-то степени они могут быть вызваны и изменениями в состоянии самих тестируемых — например болезнью, утомлением, эмоциональным напряжением, беспокойством, недавними приятными или неприятными переживаниями и т. д. Ретестовая надежность показывает, в какой степени результаты теста можно распространить на различные случаи его применения. Чем выше надежность, тем менее чувствительны тестовые показатели к случайным суточным изменениям состояния тестируемых и обстановки тестирования.

Приводя в руководстве к тесту его ретестовую надежность, всегда следует указывать, в каком интервале времени она измерена. Поскольку ретестовые корреляции постепенно снижаются по мере увеличения этого интервала, для любого теста существует не один, а бесконечное множество ретестовых коэффициентов надежности. Желательно также давать некоторые сведения о событиях, происшедших за время между двумя сеансами тестирования с теми, на ком измерялась надежность теста, и касающихся их учебы, работы, семейной жизни, консультирования, психотерапии и т. д.

Кроме желательности сообщения длины интервала между двумя тестированиями, хорошо бы знать, какими соображениями направлялся выбор именно этого интервала? Можно привести немало примеров тестов, надежность которых остается высокой в течение нескольких дней или недель, но спустя десять-пятнадцать лет их результаты уже практически не коррелируют с первоначальными. Так, многие из тестов интеллекта для дошкольников дают достаточно устойчивые показатели на протяжении дошкольного периода, но совершенно бесполезны в качестве инструментов предсказания *IQ* в позднем детстве или во взрослости. На практике, однако, чаще всего следуют простому правилу в установлении границ ретестового интервала. Обычно дисперсия ошибок тестового показателя определяется кратковременными, случайными колебаниями, происходящими в интервалах от нескольких часов до нескольких месяцев. Поэтому, при проверке этого типа тестовой надежности, стараются придерживаться небольших временных интервалов. При тестировании маленьких детей этот период должен быть еще короче, чем у испытуемых старшего возраста, поскольку в первые годы жизни связанные с возрастным развитием изменения наблюдаются ежемесячно и даже быстрее. В целом, для любого типа обследуемых лиц ретестовый интервал, по-видимому, редко превышает шесть месяцев.

Какие-либо дополнительные изменения в относительном выполнении теста одними и теми же людьми, происходящие в более длительные промежутки времени, уместнее относить к кумулятивным и прогрессирующим, а не к чисто случайным. Кроме того, такие изменения, вероятно, характеризуют более широкую сферу поведения, чем та, которая проявляется при выполнении данного теста. Так, общий уровень способности человека к обучению, пониманию технических устройств или искусства мог за 10 лет существенно измениться вследствие каких-то произошедших с ним неординарных событий. Его статус с годами мог заметно возрасти или упасть относительно статуса других людей того же возраста вследствие обстоятельств жизни дома, в школе или условий социального окружения, а также по таким причинам, как физическая болезнь или эмоциональное расстройство.

Степень влияния таких факторов на психологическое развитие человека является важной исследовательской проблемой. Однако этот вопрос не следует смешивать с вопросом надежности конкретного теста. Например, при измерении надежности те-

стов Стэнфорд—Бине, мы обычно вычисляем корреляцию между показателями, полученными с интервалом не в десять лет и даже не в один год, а в несколько недель. Конечно, с этими тестами проводились и долгосрочные ретестовые испытания, но их результаты обычно обсуждаются с точки зрения предсказуемости уровня интеллекта взрослого на основе выполнения теста в детском возрасте, а не с точки зрения надежности конкретного теста. Понятие надежности в основном ограничивается сферой краткосрочных случайных изменений, характеризующих технические характеристики самого теста, а не тестируемую область поведения.

Следует отметить, что различные поведенческие функции сами могут различаться по степени обнаруживаемых суточных колебаний. Например, на отточенности движений пальцев рук могут сказаться самые незначительные изменения в состоянии человека, никак не влияющие на понимание им речи. Если хотят получить полную оценку характера движений пальцев, свойственного конкретному человеку, то, по всей видимости, придется провести повторные тесты в течение нескольких дней; в то же время для оценки уровня его вербального понимания достаточно было бы одного сеанса тестирования. Всякий раз мы должны обращаться к анализу целей теста и всестороннему осмыслению того поведения, для предсказания которого предназначен данный тест.

Несмотря на кажущуюся простоту и очевидность методики повторного тестирования, ее применение к большинству психологических тестов представляет немалые трудности. Улучшение показателей как результат тренировки при повторении теста будет, вероятно, различным у разных людей. Кроме того, если промежуток времени между первым и вторым тестированием достаточно мал, испытуемые могут припомнить многие из своих прежних ответов. Иными словами, та же картина правильных и ошибочных ответов, вероятно, воспроизводится благодаря работе одной только памяти. Следовательно, результаты двух предъявлений теста не будут независимыми, и корреляция между ними окажется обманчиво высокой. К тому же повторное проведение может изменить саму сущность теста. В первую очередь это относится к задачам, требующим логических рассуждений или сообразительности. Испытуемый, однажды ухватив принцип решения или построив всю цепь рассуждений, в дальнейшем может воспроизводить правильный ответ, минуя промежуточные ступени. Методика повторного тестирования применима только к тем тестам, на которые их повторное проведение на одних и тех же испытуемых не оказывает заметного влияния. К этой категории относится ряд моторных тестов и тестов сенсорного различения. Однако для подавляющего большинства психологических тестов эта методика определения коэффициента надежности оказывается неприменимой.

Надежность взаимозаменяемых форм. Один из способов избежать трудностей, с которыми приходится сталкиваться при определении ретестовой надежности, — использование взаимозаменяемых форм (*alternate forms*) теста. Одних и тех же испытуемых могут тестировать в первый раз с помощью одной формы, а второй раз — с помощью другой, эквивалентной формы. Корреляция между показателями, полученными по двум формам теста, представляет его коэффициент надежности. Заметим, что такой коэффициент надежности служит мерой как временной устойчивости, так и согласованности ответов на различные выборки заданий (или формы теста). Таким образом, этот коэффициент служит смешанной характеристикой двух типов надежности. Однако поскольку оба ее типа важны для большинства целей тестирования, надежность взаимозаменяемых форм оказывается полезной мерой для оценки многих тестов.

Понятие выборочной проверки заданий, или *выборочной проверки содержания (content sampling)*[1], лежит в основе не только данного, но и других типов надежности, о которых речь пойдет дальше. Именно поэтому оно заслуживает более тщательного рассмотрения. Вероятно, каждому студенту когда-то доставались на экзамене вопросы именно по той теме, к которой он был особенно хорошо подготовлен или, напротив, знал этот материал особенно плохо. Столь знакомая всем ситуация иллюстрирует дисперсию ошибок, вызванную выборочной проверкой содержания. В какой степени показатели данного теста зависят от факторов, *специфичных* для этой конкретной подборки заданий? И если другой исследователь, работая независимо от нас, подготовил бы другой тест в соответствии с теми же требованиями, то насколько бы результаты этих тестов отличались друг от друга?

Предположим, что для оценки понимания слов общего употребления был сконструирован словарный тест, состоящий из 40 заданий. Предположим далее, что с той же целью был составлен второй тест из 40 других слов, причем были соблюдены все предосторожности, чтобы трудность теста оставалась той же самой. Различия в показателях, полученных по этим двум тестам одними и теми же людьми, иллюстрирует рассматриваемый тип дисперсии ошибок. Под действием случайных факторов, связанных с прошлым опытом разных людей, относительная трудность двух списков будет несколько меняться с переходом от одного человека к другому. Так, первый список может содержать больше слов, незнакомых испытуемому *A*, чем второй, в котором, в свою очередь, могло оказаться непропорционально много слов, незнакомых испытуемому *B*. Если оба испытуемых примерно равны по своему словарному запасу (т. е. по своим «истинным показателям»), то *B* тем не менее превзойдет *A* по первому списку, тогда как *A* превзойдет *B* по второму. Относительное положение испытуемых *A* и *B* по данным двум спискам окажется взаимно противоположным из-за случайных различий в подборке заданий.

Как и в случае ретестовой надежности, сведения о надежности взаимозаменяемых форм всегда должны сопровождаться указанием длительности временно́го интервала между двумя предъявлениями теста, а также характеристикой релевантных событий, происшедших за это время в жизни испытуемых. Если обе формы применяются *непосредственно одна за другой*, то полученная корреляция показывает только надежность параллельных форм, но ничего не говорит о надежности как временной устойчивости. Дисперсия ошибок в этом случае обусловлена колебаниями результатов при переходе от одного набора заданий к другому, а не временными флуктуациями показателей.

При разработке взаимозаменяемых форм, безусловно, следует позаботиться о том, чтобы они на самом деле были параллельными. Принципиально важно, чтобы параллельные формы конструировались как независимые тесты, отвечающие, однако, одним и тем же требованиям. Такие тесты должны содержать одинаковое число заданий, представленных в одной и той же форме и с однотипным содержанием. Диапазон и уровень трудности заданий тоже должны быть одинаковыми. Инструкции, временны́е рамки, поясняющие примеры, формат бланков и все другие аспекты теста также необходимо проверить на сопоставимость.

Следует добавить, что наличие параллельных форм желательно и по другим соображениям, помимо определения надежности теста. Взаимозаменяемые формы полез-

[1] Строгий термин *content sampling* в этом контексте можно более вольно перевести как *выборочная представленность содержания* или, короче, *выборка содержания. — Примеч. науч. ред.*

ны при повторных исследованиях и при изучении влияния некоторых промежуточных экспериментальных факторов на выполнение теста. Использование нескольких взаимозаменяемых форм служит, кроме того, средством уменьшения возможности натаскивания в выполнении тестов и обмана.

Несмотря на гораздо более широкое, сравнительно с ретестовой надежностью, применение, надежность взаимозаменяемых форм также обнаруживает ряд ограничений. Прежде всего, если изучаемые поведенческие функции подвержены значительному влиянию тренировки, использование параллельных форм ослабит, но не устранит его полностью. Конечно, если бы у всех тестируемых наблюдалось одно и то же улучшение результатов при повторном проведении теста, это не повлияло бы на корреляцию показателей, поскольку прибавление постоянной величины к каждому показателю не меняет коэффициента корреляции. Однако, скорее всего, улучшение результатов у разных людей будет неодинаковым вследствие индивидуальных различий в опыте работы с подобным материалом, в мотивации участия в тесте и по других причинам. При этих условиях эффект тренировки представляет собой еще один источник дисперсии, снижающей, в общем, корреляцию между двумя формами. Но если влияние тренированности невелико, снижение корреляции будет незначительным.

Другая проблема связана с возможным изменением сущности теста при повторном его проведении. Например, если в параллельных задачах на сообразительность применен один и тот же принцип, то большинство испытуемых, однажды найдя решение, и во второй раз применят его. В подобных случаях одной замены содержания заданий явно недостаточно для того, чтобы избежать переноса принципа принципов решения из одной формы теста на другую. Наконец, следует добавить, что для многих тестов взаимозаменяемые формы отсутствуют ввиду практических трудностей создания подлинно эквивалентных форм. В силу этих причин часто приходится обращаться к другим методам оценки надежности теста.

Надежность эквивалентных половин теста. Меру надежности можно определить и на основании однократного применения единственной формы теста, пользуясь для этого различными процедурами расщепления теста на две равноценные половины. При таком способе каждый испытуемый получает два показателя благодаря разделению теста на две эквивалентные части. Очевидно, что надежность, найденная методом расщепления, дает нам меру согласованности выборочных проверок содержания. Временная устойчивость показателей в такой характеристике надежности не представлена, поскольку она предполагает только один сеанс тестирования. Этот тип коэффициента надежности иногда называют коэффициентом внутренней согласованности, так как для его определения требуется лишь однократное проведение единственной формы теста.

Первая проблема, с которой мы сталкиваемся при применении метода расщепления, связана с тем, как разделить тест, чтобы добиться максимальной эквивалентности его половин. Всякий тест можно членить многими способами. В большинстве тестов первая и вторая половины оказались бы неэквивалентными вследствие различий в характере и уровне трудности заданий, а также в связи с кумулятивными эффектами вхождения в работу, практики, утомления, скуки и любых других факторов, воздействие которых нарастает от начала к концу теста. Подходящий для большинства целей метод состоит в вычислении показателей отдельно по четным и нечетным заданиям теста. Если задания теста были изначально расположены в порядке возрас-

тания трудности, то такое разбиение дает практически эквивалентные показатели обеих половин. Одна предосторожность, которую требуется при этом соблюдать, относится к случаю, когда тест содержит группу взаимосвязанных заданий — например, когда несколько вопросов касаются какого-то одного чертежа механического устройства в тесте технических способностей или одного и того же фрагмента текста в тесте чтения. В этом случае каждая такая группа заданий должна быть целиком отнесена либо к одной, либо к другой половине. Если задания таких групп разделить на две части, то возникнет обманчивое сходство сравниваемых показателей, так как любая ошибка в понимании задачи скажется на выполнении заданий из обеих половин.[1]

Полученные показатели по двум частям теста коррелируются обычным методом. Нужно иметь в виду, однако, что эта корреляция в действительности показывает надежность лишь половины теста. Например, если весь тест состоит из 100 заданий, то корреляция вычисляется между двумя множествами показателей, каждый из которых основан только на выполнении 50 заданий. В отличие от надежности этого типа, при расчете ретестовой надежности, как и надежности взаимозаменяемых форм, каждый показатель основывается на полном наборе заданий теста.

При прочих равных условиях, чем больше заданий содержит тест, тем выше его надежность.[2] Вполне оправданно ожидать, что чем обширнее выборка поведения, тем адекватнее и согласованнее получаемые единицы измерения. Влияние, оказываемое увеличением или сокращением теста на его коэффициент надежности, можно оценить с помощью формулы Спирмена—Брауна:

$$r_{nn} = \frac{n \cdot r_{tt}}{1 + (n-1) \cdot r_{tt}},$$

где r_{nn} — ожидаемое значение коэффициента надежности; n — отношение нового числа заданий к первоначальному; r_{tt}— полученное значение коэффициента надежности. Так, если число заданий теста возросло с 25 до 100, то $n = 4$, а если оно сократилось с 60 до 30, то $n = 1/2$. Формула Спирмена—Брауна широко используется при определении надежности методом расщепления, и во многих руководствах к тестам данные о надежности приводятся в этом виде. Применительно к расчетам надежности эквивалентных частей теста формула Спирмена—Брауна всегда предполагает удвоение числа заданий теста, и потому может быть приведена к более простому виду:

$$r_{tt} = \frac{2r_{hh}}{1 + r_{hh}},$$

где r_{hh} — корреляция эквивалентных половин теста.

Альтернативный метод вычисления надежности эквивалентных половин теста был разработан Рюлоном (Rulon, 1939). Требуется знать только дисперсию *разностей* между показателями каждого испытуемого по обеим половинам теста (SD_d^2) и дисперсию показателей по полному тесту (SD_x^2); значения этих величин подставляются в

---

[1] К настоящему времени накоплено достаточно эмпирических данных в пользу этого предположения, равно как и результатов статистического анализа таких монолитных групп заданий, или «тестов в тесте» (Sereci, Thissen, & Wainer, 1991).

[2] Увеличение числа заданий теста не влияет на временную устойчивость его показателей, а повышает только его согласованность с точки зрения выборочной проверки содержания (см. Cureton, 1965; Cureton et al., 1973).

следующую формулу, которая позволяет сразу получить характеристику надежности полного теста:

$$r_{tt} = 1 - \frac{SD_d^2}{SD_x^2}.$$

Интересно отметить связь между этой формулой и определением дисперсии ошибок. Любая разность между показателями испытуемого по двум половинам теста отражает постороннее влияние или дисперсию ошибок. Дисперсия таких разностей, поделенная на дисперсию показателей по всему тесту, дает долю дисперсии ошибок в этих показателях. Вычитая эту дисперсию ошибок из единицы, мы получаем долю «истинной» дисперсии для установленного применения теста, которая равна его коэффициенту надежности.

Надежность по Кьюдеру—Ричадсону и коэффициент альфа. Четвертый метод определения надежности, также использующий однократное предъявление единственной формы теста, основан на оценке согласованности ответов по всем заданиям теста. На эту *внутреннюю согласованность (interitem consistency — букв.* «взаимосогласованность заданий»*)* влияют два источника дисперсии ошибок: 1) выборочная представленность содержания (как в случае надежности взаимозаменяемых форм и эквивалентных половин теста) и 2) неоднородность выборочной области поведения. Чем однороднее эта область, тем выше внутренняя согласованность. Например, если один тест включает только задания на умножение, а другой — на сложение, вычитание, умножение и деление, то первый тест, вероятно, покажет более высокую внутреннюю согласованность, чем второй. Во втором, более разнородном тесте один испытуемый может лучше справиться с вычитанием, чем с другими арифметическими действиями, другой покажет относительно высокий результат в делении, но хуже проявит себя в сложении, вычитании и умножении, и т. д. Более контрастным примером однородности и разнородности мог бы служить тест, состоящий из 40 словарных заданий, и тест, содержащий 10 словарных заданий, 10 заданий на пространственные отношения, 10 — на арифметическое рассуждение и 10 — на скорость восприятия. В последнем тесте связь между выполнением различных типов заданий одним человеком может быть незначительной или полностью отсутствовать.

Очевидно, что чем однороднее тест, тем однозначнее его результаты. Предположим, что в последнем из только что упомянутых тестов из 40 заданий Смит и Джонс получили по 20 баллов. Можем ли мы заключить, что с этим тестом они справились одинаково? Вовсе нет. Смит мог правильно ответить на 10 словарных вопросов, выполнить 10 заданий на скорость восприятия и не справиться ни с одним заданием на арифметическое рассуждение и пространственные отношения. Напротив, 20 баллов Джонса могли распределиться таким образом: 5 за скорость восприятия, 5 за пространственные отношения, 10 за арифметическое рассуждение и 0 за словарь.

Суммарный показатель в 20 баллов, разумеется, можно было бы набрать путем множества других комбинаций, и тогда он имел бы совершенно иной смысл для каждой из таких различных комбинаций. С другой стороны, в более однородном словарном тесте показатель в 20 баллов, вероятно, означал бы, что испытуемый правильно указал значение примерно 20 первых слов, если задания располагались в порядке возрастания трудности. Он мог ошибиться в отношении двух-трех сравнительно легких слов, дать правильный ответ по более трудным словам, расположенным под номерами,

большими 20, но такие индивидуальные колебания ничтожны по сравнению с теми, которые обнаруживаются в более разнородном тесте.

Весьма существенным в этой связи является вопрос об относительной однородности (или неоднородности) самого критериального признака, на предсказание которого направлен тест. Хотя однородные тесты могут предпочитаться, потому что их показатели допускают довольно однозначную интерпретацию, но взятый в отдельности однородный тест, очевидно, непригоден для предсказания крайне неоднородного критериального признака. Более того, при предсказании неоднородного признака-критерия разнородность заданий теста не обязательно означала бы дисперсию ошибок. Традиционные тесты интеллекта дают хороший пример неоднородных тестов, предназначенных для предсказания неоднородного критериального признака. В подобных случаях, однако, иногда желательно составить несколько относительно однородных тестов, каждый из которых измерял бы различные аспекты неоднородного критериального признака. Тем самым однозначная интерпретация показателей теста могла бы сочетаться с адекватным охватом признака-критерия.

Самая распространенная методика оценки внутренней согласованности была разработана Кьюдером и Ричардсоном (Kuder, & Richardson, 1937). Как и в методах расщепления, внутренняя согласованность находится по данным однократного проведения единственной формы теста, но вместо использования показателей по двум эквивалентным половинам теста эта методика опирается на результаты выполнения каждого задания. Из различных формул, выведенных в указанной статье, шире других применяется так называемая формула KR — 20:

$$r_{tt} = \left(\frac{n}{n-1}\right)\frac{SD_t^2 - \sum pq}{SD_t^2}.$$

В этой формуле r_{tt} — коэффициент надежности полного теста, n — число заданий в тесте, а SD_t — стандартное отклонение суммарных показателей теста. Единственным новым элементом в этой формуле является сумма $\sum pq$ где p и q— доля испытуемых, соответственно справившихся (p) и не справившихся (q) с каждым заданием. Чтобы вычислить $\sum pq$, нужно для каждого задания найти произведение $p \times q$, а затем сложить эти произведения по всем заданиям. Поскольку в процессе конструирования теста величина p часто фиксируется для определения уровня трудности каждого задания, этот метод определения надежности требует лишь незначительных добавочных вычислений.

Можно математически доказать, что коэффициент надежности Кьюдера—Ричардсона представляет собой среднее значение коэффициентов надежности частей теста, вычисляемых для всех возможных разбиений теста надвое (Cronbach, 1951).[1] Обычный же коэффициент надежности частей теста основан на разбиении, построенном в расчете на получение эквивалентных половин. Поэтому в случае неоднородности заданий теста коэффициент Кьюдера—Ричардсона будет ниже коэффициента надежности эквивалентных половин. Следующий контрастный пример поясняет, в чем причина такого расхождения. Допустим, мы составляем тест из 50 заданий 25 различных видов (например, задания 1 и 2 — на понимание слов, 3 и 4 — на арифметическое

[1] Строго говоря, это утверждение справедливо, лишь когда коэффициенты надежности частей теста рассчитываются по формуле Рюлона (основанной на дисперсии разностей между показателями по обеим половинам теста), а не методом корреляции половин или по формуле Спирмена—Брауна (Novick & Lewis, 1967).

рассуждение, 5 и 6 — на пространственную ориентацию и т. д.). Показатели по четным и нечетным заданиям этого теста теоретически могли бы весьма тесно коррелировать друг с другом, что дало бы высокий коэффициент надежности эквивалентных половин. Но однородность этого теста была бы очень низкой в силу почти полного отсутствия согласованности результатов выполнения всех 50 заданий. В данном примере есть все основания ожидать, что коэффициент Кьюдера—Ричардсона окажется намного ниже коэффициента надежности эквивалентных половин теста. Фактически, разность между этими двумя коэффициентами может служить приблизительной числовой оценкой однородности теста.

Формула Кьюдера—Ричардсона применима лишь к тем тестам, в которых выполнение заданий оценивается как правильное либо ошибочное, или, в общем, по принципу «все или ничего». В некоторых тестах, однако, практикуется более дифференцированная форма представления результатов отдельных заданий. Например, в личностном опроснике респондент может получить различные числовые показатели по любому конкретному пункту опросника в зависимости от того, на какой из готовых категорий ответов он остановил свой выбор: «обычно», «иногда», «редко», «никогда». Для таких тестов была выведена обобщенная формула, известная как коэффициент альфа (Cronbach, 1951; Kaiser, & Michael, 1975; Novick, & Lewis, 1967). В этой формуле $\sum pq$ заменена на $\sum(SD_i^2)$ — сумму дисперсий балльных оценок по каждому заданию теста. Процедура вычислений состоит в нахождении дисперсии всех индивидуальных балльных оценок по каждому заданию с последующим суммированием этих дисперсий по всем заданиям. Полная формула коэффициента альфа выглядит следующим образом:

$$r_{tt} = \left(\frac{n}{n-1}\right)\frac{SD_t^2 - \sum(SD_i^2)}{SD_t^2}.$$

Надежность оценщика. Теперь уже очевидно, что различные типы надежности отличаются друг от друга факторами, относимыми к источникам дисперсии ошибок. В одном случае дисперсия ошибок охватывает временны́е колебания, в другом относится к различиям между наборами параллельных заданий, в третьем учитывает любую внутреннюю несогласованность теста. С другой стороны, факторы, *исключенные* из мер дисперсии ошибок, образуют два широких класса: а) факторы, чья дисперсия сохраняется в показателях, так как эти факторы составляют часть истинных различий, измеряемых тестами, и б) нерелевантные факторы, поддающиеся экспериментальному контролю. Например, в руководстве к тесту не принято сообщать об ошибках измерения, которые могут появиться в результате проведения теста в отвлекающей обстановке или в более короткое или длительное, чем это положено, время. Подобных нарушений можно избежать, и поэтому нет нужды в отдельных коэффициентах надежности, соответствующих «дисперсии отвлечения» или «дисперсии временны́х лимитов».

Большинство тестов, особенно если они предназначены для массового обследования с использованием компьютеров для вычисления показателей, настолько стандартизированы, что их проведение и регистрация результатов сводят на нет дисперсию ошибок, обусловленную этими факторами. Пользуясь такими тестами, необходимо лишь внимательно следить за выполнением соответствующих предписаний. Вместе с тем в отношении клинических тестов, применяемых при интенсивных индивидуальных обследованиях, накоплены данные о значительной *дисперсии наблюдателя (еха-*

miner variance). Благодаря использованию специальных планов эксперимента удается отделить эту дисперсию от той, которая обусловлена временными колебаниями в состоянии испытуемого или применением взаимозаменяемых форм теста.

Один источник дисперсии ошибок, который довольно легко установить, — это *дисперсия оценщика (scorer variance).* Некоторые типы тестов, — особенно тесты креативности и проективные личностные тесты, — предоставляют довольно много свободы пользователю, оценивающему ответы испытуемого и выставляющему за них определенное количество баллов. При работе с такими тестами потребность в мере надежности оценщика столь же велика, как и в более привычных коэффициентах надежности. Надежность оценщика можно определить, располагая выборкой протоколов выполнения теста, оцененного двумя специалистами независимо друг от друга. Между двумя множествами полученных таким образом показателей вычисляется обычный коэффициент корреляции, который и служит искомой мерой надежности оценщика. Если подсчет показателей теста существенно зависит от суждений пользователя, то в руководстве к тесту необходимо также привести и коэффициент надежности оценщика.

Общий обзор типов и коэффициентов надежности. Различные виды только что рассмотренных коэффициентов надежности сведены в табл. 4–2 и 4–3. В табл. 4–2 методы, применяемые для оценки каждого типа надежности, сгруппированы в зависимости от числа требуемых для этой цели форм теста и сеансов тестирования. В табл. 4–3 представлены источники дисперсии, трактуемые каждым из методов как дисперсия ошибок.

Таблица 4–2

Классификация методов измерения надежности в зависимости от требуемого числа форм теста и сеансов тестирования

Необходимое число сеансов тестирования	Необходимое число форм теста: одна	две
Один	Метод расщепления на эквивалентные половины Метод Кьюдера–Ричардсона	Метод взаимозаменяемых форм (непосредственный)
Два	Метод «тест — ретест»	Метод взаимозаменяемых форм (отсроченный)

Любой коэффициент надежности можно интерпретировать непосредственно в *процентах дисперсии показателей,* приписываемой разным источникам. Так, коэффициент надежности 0,85 означает, что 85 % дисперсии показателей теста зависят от истинной изменчивости (дисперсии) измеряемой черты, а 15 % – от дисперсии ошибок (что операционно определяется используемой расчетной процедурой). Читателю, знакомому со статистикой, напомним, что именно *квадрат* коэффициента корреляции представляет собой часть общей дисперсии. Фактически, доля истинной дисперсии в показателях теста есть квадрат корреляции между показателями, полученными по какой-то одной форме теста, и истинными показателями, свободными от случай-

ных ошибок. Эта корреляция, именуемая индексом надежности,[1] равна корню квадратному из коэффициента надежности ($\sqrt{r_{tt}}$). Если индекс надежности возвести в квадрат, то получится исходный коэффициент надежности (r_{tt}), который, следовательно, можно прямо интерпретировать как процент истинной дисперсии для указанного использования теста.

Таблица 4-3

Источники дисперсии ошибок, связываемые с коэффициентами надежности

Вид коэффициента надежности	Дисперсия ошибок
Ретестовый	Временная выборка
Взаимозаменяемых форм (непосредственный)	Выборка содержания
Взаимозаменяемых форм (с временным интервалом)	Временная выборка и выборка содержания
Эквивалентных половин теста	Выборка содержания
Кьюдера—Ричардсона и альфа	Выборка содержания и неоднородность содержания
Оценщика	Различия между оценщиками

Планы эксперимента, позволяющие получать несколько разных коэффициентов надежности на одной группе испытуемых, дают возможность проводить компонентный анализ суммарной дисперсии показателей. Рассмотрим следующий гипотетический пример. Предположим, на 100 шестиклассниках с интервалом в два месяца были проведены формы *А* и *В* теста креативности. В результате, надежность взаимозаменяемых форм составила 0,70. Кроме того, по ответам на любую из форм теста можно было вычислить коэффициент надежности эквивалентных половин.[2] Этот коэффициент, повышенный за счет применения формулы Спирмена—Брауна, составил 0,80. Наконец, надежность оценщика, полученная благодаря привлечению еще одного специалиста, проставившего новые баллы в случайно выбранных 50 протоколах ответов, оказалась равной 0,92. Анализ этих трех коэффициентов надежности с целью получения значений дисперсий ошибок показан в табл. 4-4 и на рис. 4-4. Вычитая дисперсию ошибок, связываемую только с выборкой содержания, из дисперсии ошибок, обусловленной временной выборкой и выборкой содержания, находим, что 0,10 последней можно приписать чистому влиянию временной выборки. Складывая дисперсии ошибок, связываемые с выборкой содержания (0,20), временной выборкой (0,10) и различиями между оценщиками (0,08), получаем суммарную дисперсию ошибок, равную 0,38, из чего следует, что величина истинной дисперсии равна 0,62. Эти компоненты дисперсии, выраженные в более привычной процентной форме, графически изображены на рис. 4-4. Такая классификация источников дисперсии составляет существо так называемой теории надежности как обобщаемости (*generalizability theory of reliability*). Сложные экспериментальные планы, позволяющие производить одно-

[1] Выведение индекса надежности, основанное на двух различных наборах допущений, представлено в книге Гулликсена (Gulliksen, 1950, chaps. 2 and 3).

[2] В целях более точной оценки коэффициента внутренней согласованности, корреляции между двумя половинами теста можно было вычислить для каждой формы отдельно, а затем найти среднее из двух коэффициентов корреляции, воспользовавшись подходящими статистическими методами (например, *z*-преобразованием Фишера).

Таблица 4–4

Анализ источников дисперсии ошибок в гипотетическом тесте

По надежности взаимозаменяемых форм (с временным интервалом)	1 — 0,70 = 0,30 (временная выборка + выборка содержания)
По надежности эквивалентных половин теста (формула Спирмена—Брауна)	1 — 0,80 = 0,20* (выборка содержания)
Разность	0,10* (временная выборка)
По надежности оценщика	1 — 0,92 = 0,08* (различия между оценщиками)

Суммарная оценка дисперсии ошибок* = 0,20 + 0,10 + 0,08 = 0,38*
Истинная дисперсия = 1 — 0,38 = 0,62

* Дисперсия ошибок

временную оценку большего числа источников дисперсии показателей и взаимодействий между ними, можно найти в публикациях, посвященных обстоятельной разработке этого вопроса (см., например, Brennan, 1984; Cronbach et al., 1972; Feldt, & Brennan, 1989; Shavelson, & Webb, 1991).

Надежность тестов скорости

При конструировании теста и интерпретации его показателей важно различать измерение скорости выполнения заданий и принципиальной возможности (*power*) индивидуума справиться с ними. В «чистом» *тесте скорости (speed test)* индивидуальные различия между тестируемыми полностью зависят от скорости выполнения заданий. Такой тест составляется из заданий одинаково низкой трудности, чтобы с ними заведомо могли справится все те, на кого рассчитан данный тест. Но при этом лимит времени устанавливается так, что никто не успевает выполнить всех заданий. В таких условиях показатель испытуемого отражает только скорость его работы. С другой стороны, «чистый» *тест возможностей (power test)* предоставляет достаточно времени для того, чтобы любой мог попробовать выполнить все задания. Но их трудность постепенно возрастает от задания к заданию, так что практически никто не может справиться со всеми заданиями, а значит, не может получить высший показатель.

Вообще говоря, и тесты скорости, и тесты возможностей строятся с таким расчетом, чтобы нельзя было получить высшего, предельного показателя. Такая предосто-

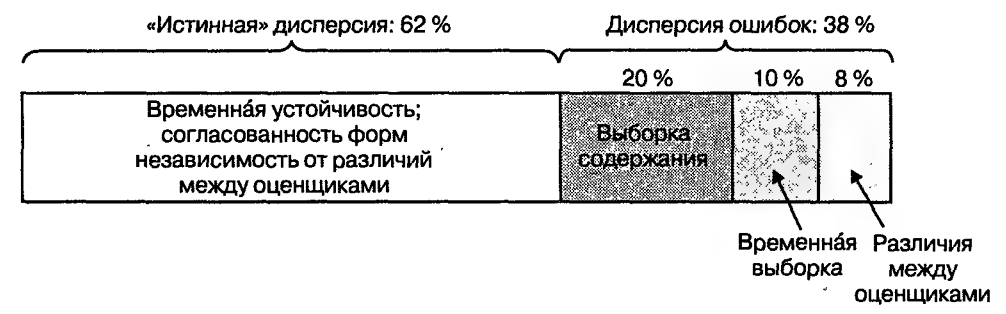


Рис. 4–4. Процентное распределение дисперсии показателя в гипотетическом тесте

рожность объясняется содержащейся в предельных показателях неопределенностью: остается неизвестным, насколько показатель конкретного человека оказался бы выше, если бы в тесте было использовано больше заданий или соответственно более трудные задания. Чтобы каждый тестируемый мог полностью продемонстрировать, на что он способен, «потолок» теста должен быть заведомо выше его возможностей либо по числу заданий, либо по уровню трудности. Исключение составляет тестирование владения предметом (или видом деятельности), как это видно на примере предметно-ориентированных тестов, обсуждавшихся в главе 3. Цель такого тестирования не в том, чтобы установить границы возможностей конкретного человека, а в определении того, достиг ли он заранее установленного уровня выполнения определенной деятельности.

На практике различие между тестами скорости и тестами возможностей — это различие в степени, и большинство тестов рассчитано на определенное соотношение скорости и возможностей. Знание этого соотношения необходимо не только для того, чтобы понять, что измеряет тот или иной тест, но и для выбора подходящих методов оценки его надежности. Коэффициенты надежности на основе однократного тестирования, наподобие тех, что определяются методами распределения заданий на четные и нечетные или по формуле Кьюдера—Ричардсона, неприменимы к тестам на скорость. Чем больше индивидуальные различия в тестовых показателях зависят от скорости выполнения, тем более завышенными оказываются коэффициенты надежности, определенные этими методами. Следующий контрастный пример поможет прояснить это утверждение. Пусть выполнение теста, состоящего из 50 заданий, полностью зависит от скорости, так что индивидуальные различия в показателе основываются исключительно на числе выполненных заданий, а не на количестве ошибок. Тогда, если испытуемый *А* получил 44 балла, он, очевидно, справился с 22 четными и 22 нечетными заданиями. Точно так же испытуемый *В* с показателем 34 балла скорее всего получил по 17 баллов за четные и нечетные задания соответственно. Следовательно, если исключить отдельные случайные ошибки, допущенные по небрежности, корреляция между показателями по четным и нечетным заданиям будет полной, т. е. равной + 1,00. Такая корреляция, однако, является ложной и не дает никакой информации о надежности теста.

Анализ методов, используемых при расчете коэффициентов надежности половин теста и Кьюдера—Ричардсона, показывает, что оба они основаны на учете согласованности *числа ошибок*, сделанных испытуемым. Если же индивидуальные различия в тестовых показателях зависят не от ошибок, а от скорости, то и в основу меры надежности должна быть положена согласованность в *скорости работы*. Когда выполнение теста зависит одновременно от скорости работы и потенциальных возможностей тестируемых, то коэффициенты надежности, вычисленные по данным однократного проведения теста, окажутся ниже 1,00, но все еще будут искусственно завышенными. Пока на индивидуальные различия в тестовых показателях существенно влияет скорость работы тестируемых, коэффициенты надежности на основе однократного тестирования не поддаются адекватной интерпретации.

Какие альтернативные методы определения надежности пригодны для тестов с выраженным скоростным компонентом? В тех случаях, когда это возможно, применяют метод повторного тестирования («тест — ретест»). С той же оговоркой применим и метод определения надежности взаимозаменяемых, эквивалентных форм. Можно воспользоваться и методом расщепления при условии, что задания теста разбиваются по временным характеристикам, а не по порядковым номерам. Иными словами, показатели по половинам теста должны основываться на раздельно нормированных по

времени частях теста. Одним из способов такого разделения является проведение двух эквивалентных половин теста с отдельно устанавливаемыми временны́ми пределами. Например, четные и нечетные задания распечатываются на разных листах и по каждому набору заданий устанавливается временно́й лимит, равный половине лимита для всего теста. Такая процедура равносильна проведению следующих друг за другом двух эквивалентных форм теста. Хотя каждая форма вдвое короче целого теста, показатели тестируемых, как обычно, основываются на результатах выполнения всего теста. По этой причине, чтобы определить надежность полного теста, нужно воспользоваться формулой Спирмена—Брауна или другой подходящей для такого случая формулой.

Если раздельное проведение двух половин теста невозможно, то вместо этого можно воспользоваться разделением полного времени теста на четыре части с регистрацией результатов отдельно для каждой четверти. Это легко осуществить, прося испытуемых по условленному сигналу проводящего тест отметить крестиком выполняемое в данный момент задание. Число заданий, правильно выполненных за первую и четвертую части полного временно́го лимита, можно затем объединить для вычисления показателя по первой половине теста. Показатель по другой половине теста будет тогда соответствовать числу заданий, с которыми испытуемый справился за вторую и третью четверти. Такая комбинация четвертей способствует нейтрализации кумулятивных эффектов тренировки, утомления и других факторов. Этот метод особенно хорошо работает, когда задания не отличаются резко друг от друга по уровню трудности.

В каких случаях скоростной компонент следует считать существенным? При каких условиях нужно соблюдать рассмотренные выше меры предосторожности? Очевидно, само по себе использование лимита времени еще не означает, что мы имеем дело с тестом скорости. Если все тестируемые укладываются в отведенное время, то скорость работы не сказывается на показателях. В качестве грубой числовой характеристики выраженности скоростного компонента, казалось бы, можно взять процент тестируемых, не успевающих закончить тест в установленное время. Однако даже если никто не укладывается в отведенные временны́е рамки, скорость выполнения может оказаться тут ни при чем. Например, если все тестируемые выполнят 40 заданий из 50, то индивидуальные различия в скорости отсутствуют, хотя никто не успевает выполнить весь тест.

Существенным здесь оказывается следующий вопрос: «В какой степени индивидуальные различия в тестовых показателях определяются скоростью работы?» Выражаясь более специальным языком, нам нужно знать, какую долю суммарной дисперсии тестовых показателей составляет дисперсия скорости. Эту долю можно приблизительно оценить, вычислив дисперсию числа выполненных разными испытуемыми заданий и разделив ее на суммарную дисперсию тестовых показателей $\left(SD_c^2 / SD_t^2 \right)$. Для только что приводившегося примера, когда все испытуемые выполнили по 40 заданий, числитель этой дроби равен нулю, поскольку отсутствуют индивидуальные различия в числе выполненных заданий $\left(SD_c^2 = 0\right)$. Таким образом, в чистом тесте возможностей данный индекс будет равен нулю. Напротив, если суммарная дисперсия теста $\left(SD_t^2\right)$ определяется индивидуальными различиями в скорости, то обе дисперсии будут равны и их отношение обратится в 1,00. Для определения этой доли дисперсии в суммарной дисперсии тестовых показателей разработан ряд более точных методов, но их детальное обсуждение выходит за рамки настоящей книги.

Пример влияния скорости работы на коэффициенты надежности, определяемые по результатам однократного проведения теста, дают данные, собранные в исследовании первой редакции *SRA* Тестов первичных умственных способностей для возраста 11–17 лет (Anastasi, & Drake, 1954). В этой работе надежность каждого теста сначала определялась обычным методом расщепления теста на четные и нечетные задания. Соответствующие коэффициенты приведены в первой строке табл. 4–5. Затем вычислялись коэффициенты надежности на основе корреляции показателей по половинам, путем разделения каждого теста на две части с отдельно устанавливаемыми лимитами времени. Эти коэффициенты приведены во второй строке табл. 4–5. Вычисление «скоростных индексов» показало, что тест «вербальное понимание» оказался, по существу, тестом возможностей, тогда как тест «логическое рассуждение» в несколько большей степени зависел от скорости работы. Из табл. 4–5 видно, что при выборе адекватного метода оценки надежности, коэффициент надежности для теста «пространственные отношения» составил 0,75 против искусственно завышенного коэффициента 0,90, полученного методом расщепления теста на четные и нечетные задания. Аналогично этому, надежность теста «логическое рассуждение» упала с 0,96 до 0,87, а «числового» теста — с 0,92 до 0,83. С другой стороны, вычисленные этими двумя методами коэффициенты надежности для теста «вербальное понимание», содержащего лишь минимальный скоростной компонент, обнаруживают незначительное различие.

Таблица 4–5

Коэффициенты надежности четырех тестов, входящих в *SRA* Тесты первичных умственных способностей для возраста 11–17 лет (Первая редакция)

Коэффициент надежности, определяемый:	Вербальное понимание	Логическое рассуждение	Пространственные отношения	Числовой
методом разделения заданий на четные и нечетные (в одном сеансе тестирования)	0,94	0,96	0,90	0,92
методом установления отдельных временных лимитов для половин теста	0,90	0,87	0,75	0,83

(По данным из Anastasi, & Drake, 1954)

Зависимость коэффициентов надежности от обследуемой выборки

Изменчивость. Важным условием, влияющим на величину коэффициента надежности, является характер группы, используемой для измерения надежности теста. В первую очередь на любой коэффициент корреляции влияет диапазон индивидуальных различий в группе. Если, к примеру, владение орфографией у всех членов группы находится примерно на одном уровне, то в этой группе корреляция орфографической способности с любыми другими способностями будет близка к нулю. Иначе говоря,

внутри такой группы невозможно было бы предсказать относительное положение индивидуума по какой-либо способности на основе знания его показателя по орфографическому тесту.

Другим, менее контрастным примером может служить корреляция между двумя тестами способностей, такими как тест вербального понимания и тест арифметического рассуждения. Если эти тесты провести в достаточно однородной группе, скажем, среди 300 студентов второго курса, то корреляция между соответствующими показателями, вероятно, будет очень низкой. Вследствие ограничения диапазона изменчивости внутри такой «отборной» выборки студентов колледжа вряд ли удастся обнаружить какую-либо связь между вербальной способностью и способностью к рассуждению с числами у ее представителей. С другой стороны, проведи мы те же тесты на неоднородной выборке из 300 человек — от умственно отсталых до выпускников колледжей, — результатом наверняка будет высокая корреляция между их показателями. Умственно отсталые по *обоим* тестам получат более низкие показатели, чем лица с высшим образованием, и подобное соотношение показателей сохранится в других подгруппах внутри этой крайне неоднородной выборки.

Анализ гипотетической диаграммы рассеяния на рис. 4–5 послужит дополнительной иллюстрацией зависимости коэффициентов корреляции от диапазона изменчивости, или степени индивидуальных различий внутри группы. Диаграмма показывает высокую положительную корреляцию в полной, неоднородной группе, так как показатели тесно группируются вдоль диагонали, идущей от левого нижнего к правому верхнему углу. Если теперь рассмотреть только подгруппу, попадающую в небольшой прямоугольник в правой верхней части диаграммы, с первого взгляда видно, что корреляция между двумя переменными в этой подгруппе близка к нулю. Испытуемые, попадающие в выделенную прямоугольником ограниченную область значений обеих переменных, представляют собой весьма однородную группу, наподобие упомянутой выше группы второкурсников.

Как и все коэффициенты корреляции, коэффициенты надежности зависят от изменчивости выборки, на которой они определяются. Так, если коэффициент надежности, приводимый в руководстве к тесту, был определен на группе учеников 4–12-х классов, то нельзя полагать, что коэффициент надежности будет столь же высоким, скажем, в выборке восьмиклассников. Когда мы собираемся использовать тест для выявления индивидуальных различий в пределах более однородной выборки, чем группа стандартизации, коэффициент надежности следует заново определить именно на такой выборке. В элементарных учебниках по статистике приводятся формулы для расчета ожидаемого значения коэффициента надежности при увеличении или уменьшении стандартного отклонения показателей определенной группы. Однако предпочтительней пользоваться коэффициентами надежности, вычисленными эмпирически на группе, сравнимой с той, в которой предполагается использовать тест. Для тестов, охватывающих широкий диапазон возраста или способности, в руководствах должны приводиться отдельные коэффициенты надежности для относительно однородных подгрупп внутри выборки стандартизации.

Уровень способности. Коэффициент надежности изменяется не только в зависимости от степени индивидуальных различий в выборке, но его величина может также разниться в группах, различающихся средним уровнем измеряемой способности. Влияние последнего фактора обычно нельзя предсказать или оценить, пользуясь ста-

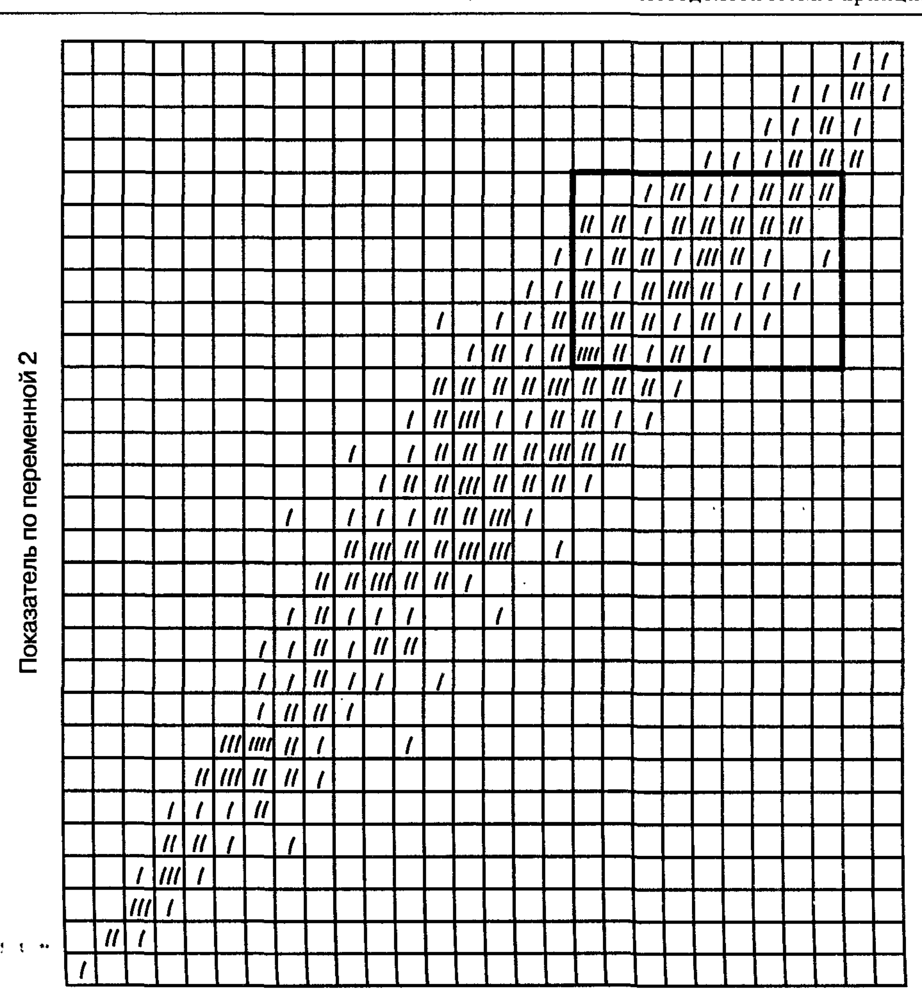


Рис. 4–5. Влияние ограниченного диапазона изменчивости переменных на величину коэффициента корреляции

тистическими методами. Это влияние может быть определено только эмпирической проверкой теста в группах, отличающихся возрастом или уровнем способности. Такие различия в надежности одного теста могут, отчасти, являться результатом того, что на каждом уровне трудности теста измеряется несколько иное сочетание способностей. Как, впрочем, могут вызываться и в результате изменения длины теста (количества заданий) на разных возрастных уровнях. Даже когда число предлагаемых заданий одинаково, верхний и нижний края теста часто не обеспечивают на соответствующем уровне трудности достаточного количества заданий, чтобы дать возможность испытуемым в полной мере продемонстрировать то, на что они способны (эффекты «потолка» или «пола»). В других тестах надежность может быть относительно низкой для младших и менее способных групп, так как в данном случае на показателях сильно сказывается стремление тестируемых угадать ответ.

Очевидно, что каждый коэффициент надежности следует сопровождать полной типологической характеристикой группы, на которой он определялся. Особое внимание следует уделять изменчивости и уровню изучаемой способности в выборке. Приводимый коэффициент надежности применим только к выборкам, сходным с теми, на которых он был определен. В настоящее время при конструировании тестов все чаще применяется разбиение выборки стандартизации на более однородные подгруппы по признаку возраста, пола, года обучения, рода занятий и т. п., причем для каждой такой подгруппы приводятся свои коэффициенты надежности. При таких условиях коэффициенты надежности более соответствует тем выборкам, в которых тест будет применяться на практике.

Стандартная ошибка измерения

Интерпретация индивидуальных показателей. Надежность теста можно выразить через стандартную ошибку измерения (*SEM* — сокр. от *standard error of measurement*), называемую также стандартной ошибкой показателя. Эта мера особенно удобна для интерпретации индивидуальных показателей. Поэтому для многих целей тестирования она более полезна, чем коэффициент надежности. Зная коэффициент надежности теста, стандартную ошибку измерения легко вычислить по следующей формуле:

$$SEM = SD_t\sqrt{1-r_{tt}},$$

где SD_t — стандартное отклонение показателей теста; r_{tt} — коэффициент надежности, оба вычисленные на одной группе. Например, если стандартные показатели *IQ* по конкретному тесту интеллекта имеют $SD_t = 15$ и коэффициент надежности $r_{tt} = 0{,}89$, то SEM_{IQ} в этом тесте равна $15\sqrt{1-0{,}89} = 15\sqrt{0{,}11} \approx 15 \cdot 0{,}33 \approx 5$.

Чтобы понять, о чем нам говорит стандартная ошибка показателя, предположим, что мы располагаем сотней стандартных *IQ*, полученных единственным ребенком, Жанет, по упомянутому выше тесту интеллекта. Вследствие разного рода случайных ошибок, обсуждавшихся в данной главе, эти показатели будут варьировать вокруг истинного показателя Жанет, подчиняясь нормальному распределению. Среднее этого распределения ста показателей можно принять за «истинный показатель» для данного использования теста, а стандартное отклонение — за соответствующую *SEM*. Как и любое стандартное отклонение, стандартную ошибку можно интерпретировать в единицах плотности нормального распределения (см. главу 3, рис. 3–3). Напомним, что при нормальном распределении в интервал $M \pm 1\sigma$ попадает приблизительно 68 % всех случаев. Следовательно, имеется примерно 2 шанса против 1 (точнее, 68 : 32), что *IQ* Жанет по этому тесту будут колебаться в пределах ± 1 *SEM* или 5 единиц в обе стороны от ее истинного *IQ*. Если ее истинный *IQ* = 110, можно ожидать, что в 2/3 (68 %) случаев показанные ею результаты попадут в интервал между 105 и 115.

Когда мы хотим чувствовать себя увереннее в наших предсказаниях, мы можем выбрать более высокое соотношение шансов, чем 2 : 1. Из рис. 3–3 (глава 3) видно, что интервал $M \pm 3\sigma$ покрывает 99,7 % случаев. Обратившись к таблицам плотности нормального распределения, можно удостовериться, что интервал $M \pm 2{,}58\sigma$ включает точно 99 % случаев. Следовательно, имеется 99 шансов против 1, что *IQ* Жанет попадет в интервал с границами, отстоящими на 2,58 *SEM* или на 2,58 × 5 = 13 единиц в обе

стороны от ее истинного *IQ*. Таким образом, можно с 99 % степенью уверенности (1 шанс ошибиться против 100) утверждать, что *IQ* Жанет при любом одиночном проведении этого теста будет лежать в пределах значений от 97 до 123 (100 — 13 и 110 + + 13). Если бы Жанет предъявили 100 эквивалентных тестов, то ее *IQ* мог бы выйти за границы этой области значений только однажды.

Разумеется, на практике мы не располагаем истинными показателями; обычно в нашем распоряжении имеются лишь показатели, полученные при одном-единственном проведении теста. В этих обстоятельствах мы можем применить выше приведенные рассуждения в обратном порядке. Если маловероятно, что полученный тестируемым показатель отклонится от его истинного показателя более чем на 2,58 *SEM*, мы могли бы утверждать, что этот *истинный* показатель должен лежать в пределах 2,58 *SEM* от *полученного* им показателя. Хотя нельзя установить вероятность справедливости этого утверждения для любого отдельного показателя, полученного конкретным испытуемым, можно сказать, что оно будет верным для 99 % всех возможных случаев. Следуя этому рассуждению, Галликсен (Gulliksen, 1950, p. 17–20) предложил использовать стандартную ошибку измерения для оценки разумных границ истинного показателя у лиц с любым полученным в единичном измерении показателем. В психологическом тестировании стало обычным интерпретировать ошибку измерения именно с точки зрения таких «разумных границ», и в этой книге она тоже будет интерпретироваться с этих позиций.[1]

Стандартная ошибка измерения и коэффициент надежности — это явно взаимозаменяемые способы выражения надежности теста. В отличие от коэффициента надежности ошибка измерения не зависит от изменчивости внутри группы, на которой она вычисляется. Выражаясь в единицах индивидуальных показателей, она не меняется в зависимости от того, проводятся ли измерения в однородной или неоднородной группе. С другой стороны, приводимая в единицах показателя, ошибка измерения не допускает прямого сравнения при переходе от теста к тесту. Обычные проблемы сопоставимости единиц возникают всякий раз, когда ошибка измерения сообщается в виде числа арифметических задач, количества слов словарного теста и т. п. Отсюда, если мы хотим сравнить надежность *различных тестов*, лучше пользоваться коэффициентами надежности. Однако для интерпретации *индивидуальных показателей* более подходит стандартная ошибка измерения.

Но как в отношении коэффициентов надежности, так и в отношении ошибок измерения нельзя предположить, что они остаются постоянными при изменении *уровня способности* в широком диапазоне. Обсуждаемые в предыдущем разделе различия в коэффициентах надежности сохраняются в тех случаях, когда ошибки измерения вычисляются для разных уровней одного и того же теста. Полное решение этой проблемы обеспечивается *IRT* методами анализа заданий, упоминавшимися в главе 3. Покрывая широкий диапазон тестируемой способности, эти методы позволяют выразить точность измерения теста в виде функции уровня такой способности. Метод *IRT*

---

[1] Предлагались и другие методы, использующие ожидаемое значение «истинного» показателя в качестве центра доверительного интервала (Dudek, 1979; Glutting, McDermott, & Stanley, 1987). При высоком коэффициенте надежности этот метод малоэффективен; когда же он низок, то и истинный показатель, и величину доверительного интервала удается рассчитать по столь же ненадежному коэффициенту надежности. Более того, можно выбрать оптимальный метод в зависимости от конкретной цели предполагаемого использования тестовых показателей (например, для долгосрочного прогноза или оценки текущих результатов).

позволяет получить информационную, или *характеристическую кривую теста (test information curve)*, зависящую только от включенных в данный тест заданий и дающую оценку ошибки измерения для каждого уровня способности. Более обстоятельно эти методы рассматриваются в главе 7.

Стандартная ошибка измерения (или какая-то другая числовая характеристика точности измерения) предохраняет от придания чрезмерного значения одному-единственному числовому показателю. Это применение *SEM* настолько важно, что все больше публикуемых в настоящее время тестов сопровождается информацией о показателях, но не в виде отдельных чисел, а в форме интервала показателей, внутри которого, вероятно, находится истинный показатель каждого конкретного индивидуума. Совет колледжей приводит данные о *SEM* и разъясняет, как ими пользоваться, не только в материалах, распространяемых среди консультантов в школах и колледжах, но и в индивидуальных заключениях по результатам *SAT*, рассылаемых прошедшим тестирование. *SEM* также включается в инструктивные материалы для того, чтобы учащиеся могли сориентироваться в отношении набранных ими тестовых баллов. Информация о стандартных ошибках измерения обеспечивается и при интерпретации результатов Письменных экзаменов для аспирантов (*GRE 1995–1996 guide*).

Интерпретация различий в показателях. Особенно важно учитывать надежность теста и ошибки измерения в тех случаях, когда оценивают *различия* между двумя показателями. Мышление, опирающееся на понятие интервала значений, которые каждый показатель может принимать в зависимости от действия случайных факторов, предостерегает против придания чрезмерного значения небольшим различиям в показателях. Подобную осторожность желательно проявлять как при сравнении показателей теста у разных людей, так и при сравнении показателей различных способностей одного человека. Аналогично этому, изменения показателей вследствие обучения или воздействия других экспериментальных переменных нужно интерпретировать с учетом ошибок измерения.

Часто возникающий по поводу тестовых показателей вопрос касается относительного положения человека в различных областях поведения и деятельности. Действительно ли у Дорис вербальные способности более выражены, чем арифметические? Есть ли основания считать, что Том более способен к работе с техникой, нежели со словом? Если при использовании одной из батарей тестов способностей Дорис получила более высокий показатель по вербальному, чем по числовому субтесту, а Том набрал больше баллов по механическому, чем по вербальному субтесту, то с какой уверенностью можно утверждать, что они могли бы иметь те же показатели при повторном тестировании с другой формой батареи? Иными словами, не могут ли полученные различия в показателях быть всего лишь результатом случайного отбора конкретных заданий в данных субтестах — вербальном, математическом и механическом? Подобные вопросы особенно важны для правильной интерпретации показателей по универсальным тестовым батареям способностей и черт личности (Anastasi, 1985a). Примеры и более подробное обсуждение проблем, которые нужно учитывать при интерпретировании индивидуального профиля показателей по таким батареям, можно найти в главах 8 и 9 (для тестов способностей) и главе 13 (для тестов личности).

В связи с растущим интересом к интерпретации профилей показателей издатели тестов разработали формы бланков, позволяющие давать оценку показателей в единицах их ошибок измерения. Примером может служить форма регистрации индиви-

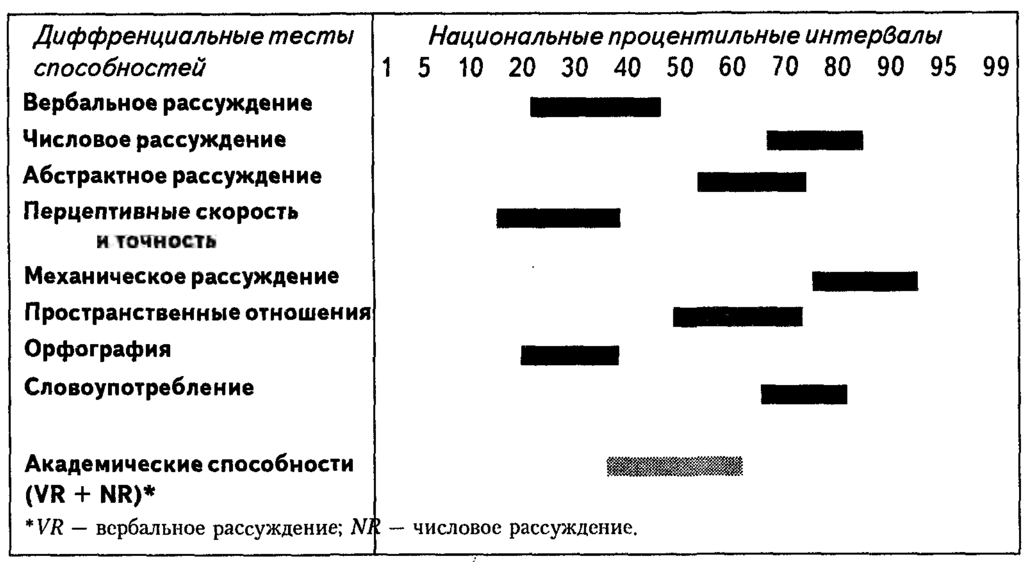


Рис. 4–6. Профиль показателей по Дифференциальным тестам способностей, построенный с использованием процентильных интервалов.
(По данным из Individual Report, Differential Aptitude Tests, 5th ed. Copyright © 1990 by The Psychological Corporation. Воспроизведено с разрешения)

дуальных показателей для использования с Дифференциальными тестами способностей (*DAT*), позволяющая представлять информацию в том виде, как показано на рис. 4–6. На этом бланке процентильные показатели по каждому субтесту батареи изображены в виде процентильных интервалов — полосок с фактическим процентильным показателем в центре. Длина каждой такой процентильной полоски соответствует 2 *SEM*, по 1 *SEM* в обе стороны от фактического показателя. Следовательно, вероятность того, что «истинный» показатель индивидуума заключен внутри представленного этой полоской интервала, выражается соотношением шансов 2 : 1 (или 68 : 32). При интерпретации профилей пользователям теста рекомендуется не придавать значения различиям между показателями, чьи процентильные интервалы перекрывают друг друга, особенно если перекрытие превышает половину их длины. В профиле, приведенном на рис. 4–6, например, различие между показателями словесного и числового рассуждения, по-видимому, отражает подлинную разницу в уровне способности, чего, вероятно, нельзя сказать о различии в показателях числового и абстрактного рассуждения. Различие же между показателями абстрактного и механического рассуждения попадает в зону неопределенности.

Неплохо запомнить, что стандартная ошибка разности (двух) показателей больше ошибки измерения каждого из них в отдельности. Это вытекает из того, что на величину этой разности влияют случайные ошибки, присутствующие в *обоих* показателях. Зная стандартные ошибки измерения показателей, стандартную ошибку разности можно вычислить по следующей формуле:[1]

[1] Эту формулу не следует путать с формулой для вычисления стандартной ошибки разности *выборочных средних*, которая включает в качестве члена коэффициент корреляции в тех случаях, когда две сравниваемые переменные являются зависимыми. *Ошибки измерения* двух переменных — это случайные ошибки и, следовательно, независимы по предположению.

$$SE_{diff.} = \sqrt{(SEM_1)^2 + (SEM_2)^2},$$

где $SE_{diff.}$ — стандартная ошибка разности показателей, а SEM_1 и SEM_2 — стандартные ошибки измерения отдельных показателей. Заменяя SEM_1 и SEM_2 на $SD\sqrt{1-r_{11}}$ и $SD\sqrt{1-r_{22}}$ соответственно, можно выразить $SE_{diff.}$ через коэффициенты надежности:

$$SE_{diff.} = SD\sqrt{2 - r_{11} - r_{22}},$$

здесь SD — стандартное отклонение, одинаковое для тестов 1 и 2, так как показатели по ним должны быть выражены в единицах одной шкалы, чтобы их можно было сравнивать.

Можно проиллюстрировать применение этой формулы на примере вербального и невербального *IQ* пересмотренной шкалы интеллекта Векслера для взрослых (*WAIS-R*). Найденная методом расщепления надежность этих показателей равна соответственно 0,97 и 0,93. Стандартные *IQ WIAS-R* имеют шкалу со средним $M = 100$ и $SD = 15$. По этим данным можно вычислить стандартную ошибку разности между этими двумя показателями:

$$SE_{diff.} = 15\sqrt{2 - 0{,}97 - 0{,}93} \approx 4{,}74.$$

Чтобы определить максимальную величину разности между показателями, которую можно получить в силу действия чисто случайных факторов, например на уровне значимости 0,05, умножим стандартную ошибку разности 4,95 на 1,96, что даст 9,7, т. е. приблизительно 10 единиц шкалы. Следовательно, различия между вербальным и невербальным *IQ WIAS-R* у любого тестируемого должны быть не меньше 10 единиц, чтобы их можно было считать значимыми на уровне 0,05.[1]

Оценка надежности в тестировании владения предметом и критические показатели

В главе 3 речь шла о том, что предметно-ориентированные тесты обычно (хотя и не всегда) оценивают выполнение с точки зрения совершенного владения (мастерства, квалификации), а не степени достижения. Статистическим следствием этого является снижение вариативности (изменчивости) показателей тестируемых. Теоретически, если обучение каждого индивидуума продолжать до полного овладения конкретным навыком или умением, вариативность упадет до нуля. В одном из предыдущих разделов этой главы объяснялось, что *любая* корреляция, и коэффициент надежности в том числе, зависит от диапазона изменчивости результатов в группе, на которой она вычисляется. С уменьшением вариативности выборочных данных падает и величина коэффициента корреляции. Следовательно, было бы неправильно оценивать надежность большинства предметно-ориентированных тестов обычными методами, применяя их к группе лиц уже после того, как они достигли заранее установленного уровня владения знаниями, умениями и навыками. При этих условиях даже тесты с высокой

[1] Более точные оценки можно получить при использовании фактических коэффициентов надежности и стандартных отклонений, рассчитываемых в каждой возрастной группе. В этом случае минимальные значимые различия между вербальным и невербальным IQ на 5 %-ном уровне, согласно руководству по *WIAS-R*, колеблются от 8,83 до 12,04. Тем не менее большая их часть близка к 10 единицам.

временно́й устойчивостью и внутренней согласованностью могли бы дать коэффициент надежности, близкий к нулю.

Это кажущееся препятствие на пути оценивания надежности таких тестов появляется тогда, когда упускают из виду специфическое назначение предметно-ориентированных тестов. Фактически, эти тесты используют, в основном, для различения тех, кто уже приобрел требуемые для определенной деятельности знания, умения и навыки, от тех, кому это пока не удалось сделать. Конкретные цели проведения таких тестов могут широко варьироваться — от выдачи водительских прав или назначения на должность до перехода на следующую ступень в программе индивидуального обучения или зачисления на определенный университетский курс. Тем не менее во всех таких ситуациях сам факт использования теста предполагает ожидание вариативности результатов его выполнения. Значительная доля этой вариативности отражает индивидуальные различия в результатах предшествующего обучения выполнению соответствующих функций.

Специально для оценки надежности предметно-ориентированных тестов было разработано больше дюжины различных методов (Berk, 1984b; Brennan, 1984; Subkoviak, 1984). Некоторые из этих методов подходят для простых решений типа «владеет/не владеет», при которых все ошибки классификации считаются в равной степени серьезными, независимо от того, насколько они отклоняются от критического показателя. В таких условиях можно провести тест и ретест с параллельными формами, чтобы найти процент лиц, для которых одинаковое решение принимается в обоих случаях. Эти данные можно подвергнуть дальнейшему анализу, вычисляя коэффициенты согласия и определяя их уровни значимости. Другие методы учитывают фактические показатели по двум тестированиям, и позволяют получить числовые характеристики, отражающие отклонение каждого индивидуального показателя в ту или иную сторону от любого заданного значения критического показателя. При выборе конкретного метода следует принимать в расчет характер и области применения теста, положение критических показателей и другие психометрические характеристики используемого теста. Соответствующие соображения широко рассмотрены в специальной литературе (см. Berk, 1984a; Feldt, & Brennan, 1989).

5 ВАЛИДНОСТЬ: ОСНОВНЫЕ ПОНЯТИЯ

Валидность теста — понятие, относящееся к тому, *что* тест измеряет и *насколько хорошо* он это делает. Валидность любого теста говорит нам о том, какие выводы можно сделать из полученных по нему показателей. В этой связи следует предостеречь от принятия названия теста за отличительный признак того, что им измеряется. Названия тестов выполняют функцию коротких, удобных опознавательных признаков, и только. По большей части эти названия слишком широки и расплывчаты, чтобы по ним можно было установить, к какой именно области поведения относится тот или иной тест. Правда, в последнее время наметилась тенденция давать тестам более конкретные и эмпирически обоснованные названия. Установить, какое свойство измеряет данный тест, можно лишь на основе изучения объективной информации и эмпирических операций, применявшихся при установления его валидности. Да и сами сведения о валидности теста невозможно представить в общих чертах. Ни о каком тесте нельзя сказать, что он имеет «высокую» или «низкую» валидность вообще. Его валидность должна устанавливаться в отношении того конкретного применения, ради которого он выбирается.

В принципе, все методы определения валидности теста имеют дело с тем, как выполнение теста соотносится с другими независимо наблюдаемыми фактами исследуемых характеристик поведения. Существуют многочисленные методы исследования подобных соотношений, описанные к тому же под различными названиями. Их традиционные названия отражают разные аспекты валидности, равно как и особый интерес к отдельным областям применения тестов. Вместе с развитием тестов и расширением сферы их применения видоизменялись и понятия валидности (Anastasi, 1986a; Messick, 1988, 1989).

Развитие понятий валидности теста

К самым истокам тестирования восходит применение тестов для оценки усвоенного людьми содержания в конкретных областях знаний или деятельности. В наши дни это применение тестов представлено переводными и выпускными экзаменами в

школе и тестами для получения водительских прав или права занимать определенную должность. Этот тип теста, обычно определяемый как тест достижения, принято оценивать путем сравнения его содержания с содержанием той области, для оценки которой он предназначается. Такой дескриптивный (описательный) подход до сих пор сохраняет своё значение в том, что касается валидизации тестов, и будет рассмотрен в первом разделе этой главы.

С переходом тестирования во вторую фазу своего развития, главный интерес переместился с констатации на предсказание. Как разные люди будут реагировать на данную ситуацию сейчас или через какое-то время? Какой будет эта индивидуальная реакция в разных точно установленных ситуациях? Действие (или деятельность) в той ситуации, для которой хотели предсказать поведение, стали назвать критерием. Соответственно, валидность теста обычно сообщалась в виде коэффициента корреляции между показателями теста и прямой, независимой мерой такого критерия. Этот метод особенно подходит для тестов, применяемых при отборе или распределении лиц, поступающих в учебные заведения, на работу или желающих пройти определенный курс лечения. Так, для теста механических способностей критерием могла бы быть эффективность последующей работы в должности механика, для теста академических способностей — оценки в колледже, а для шкалы нейротизма — оценки товарищей или другие доступные сведения о поведении человека в различных жизненных ситуациях.

Современный этап в истории тестирования отражает две главные тенденции: 1) усилившуюся теоретическую ориентацию и 2) тесное сцепление психологической теории с верификацией посредством эмпирической и экспериментальной проверки гипотез. Эти тенденции носят явный характер в конструировании и валидизации тестов, как, впрочем, и в других областях психологии как науки в целом (Anastasi, 1992a, 1992b, 1995).

Один из результатов этих тенденций — растущее признание ценности конструктов в том, что касается описания и понимания поведения человека. Конструкты — это широкие категории, выводимые логическим путем из общих признаков, свойств или черт, обнаруживающих себя в непосредственно наблюдаемых поведенческих переменных. Сами же конструкты, будучи теоретическими категориями, недоступны непосредственному наблюдению.

Интерес к конструктам привел к введению нового понятия, которое сначала считалось еще одной, третьей, разновидностью понятия валидности теста, именно конструктной валидности (AERA, APA, NCME, 1985; APA, AERA, NCME, 1974; Cronbach, & Meehl, 1955). Со временем конструктную валидность признали в качестве основного, базисного понятия валидности, включающего все ее остальные виды, поскольку именно она точно определяет, что измеряется данным тестом. Методы установления содержательной и прогностической валидности относятся к разряду тех многих средств получения информации, которые способствуют более точному определению и пониманию конструктов, оцениваемых тестами. В то же время эти методы дают информацию, представляющую самостоятельную ценность, и сохраняют свое первостепенное значение при оценке применяемых в ряде областей тестов. А потому понятия (и соответствующие термины) содержательной и прогностической валидности остались в употреблении, несмотря на их интеграцию в единое понятие конструктной валидности.

Методы описания содержания

Сущность. Методы установления валидности через описание содержания, по существу, заключаются в систематической проверке содержания теста на соответствие репрезентативной выборке измеряемой области поведения. Такая процедура валидизации обычно применяется к тестам, предназначенным для измерения того, насколько человек овладел конкретными навыками или учебным предметом. Может создаться впечатление, что для установления валидности любого такого теста достаточно было бы простого просмотра его содержания. Например, тест на умножение, правописание или бухгалтерские навыки, казалось бы, должен быть валидным по определению, если состоит из заданий на умножение, правописание или ведение бухгалтерских операций соответственно.

Решение, однако, не столь просто, как это может показаться. Сразу же возникает проблема формирования выборки заданий, адекватно отражающих всю оцениваемую предметную область. Поэтому тестируемая область поведения сначала должна быть подвергнута систематическому анализу, с тем чтобы существовала уверенность в полном и пропорциональном охвате ее главных аспектов заданиями теста. Например, тест можно легко перегрузить теми аспектами исследуемой области, по которым проще составить объективные задания. Поэтому рассматриваемую предметную область следует описывать заранее, и как можно полнее, а не определять после того, как тест уже составлен. Правильно построенные образовательные тесты должны охватывать цели обучения, а не только его конкретные темы. Содержание, следовательно, необходимо определять достаточно широко, включая в него помимо знания фактического материала такие важнейшие цели обучения, как применение изученных правил и объяснение фактов. Кроме того, валидность больше зависит от релевантности тестовых ответов индивидуума рассматриваемой сфере поведения, чем от очевидной релевантности содержания тестовых заданий. Простая проверка содержания теста может и не выявить те процессы, которые действительно обеспечивают выполнение теста испытуемыми.

Важно также избежать неоправданных обобщений в отношении области поведения, выборочно проверяемой тестом. Если, например, орфографический тест с множественным выбором ответов измеряет способность распознавать правильно и неправильно написанные слова, то из этого не следует, что он также измеряет способность правильно написать диктант, частоту орфографических ошибок в сочинении и другие аспекты умения писать без орфографических ошибок (Ahlström, 1964; Knoell, & Harris, 1952). Еще одна трудность возникает в связи с возможным влиянием посторонних факторов на показатели теста. Например, на результаты экзаменационного теста по математике или механике может чрезмерно повлиять способность понимать словесные инструкции или скорость выполнения простых, стандартных задач.

Конкретные методы. Содержательная валидность теста обеспечивается с самого начала благодаря отбору соответствующих заданий. Что касается образовательных тестов, подготовке их заданий предшествует полный систематический просмотр соответствующих учебников и учебных программ, а также консультации со специалистами по данному предмету. На основе собранной таким путем информации составля-

ется *спецификация теста* (*test specifications*)[1] для составителей заданий. В ней указывается охватываемые тестом области содержания или темы, проверяемые учебные цели-задачи или способы действия, а также относительное значение отдельных тем и способов. В заключение должно быть указано требуемое число заданий каждого типа по каждой теме. Например, тест для оценки умения читать может включать понимание лексики в контексте, дословное понимание содержания и умение делать правильные выводы из приведенной информации. Кроме того, он может предполагать выборочную проверку материала из разных источников, таких как рассказы, стихи, газетные статьи или инструкции по эксплуатации оборудования. Тест по математике может охватывать вычислительные навыки, решение словесно сформулированных задач и применение усвоенных способов решения в новых и непривычных условиях.

Данные о содержательной валидности, приводимые в руководстве к тесту учебных достижений, должны сопровождаться описанием тех методов, которыми обеспечивались целесообразный отбор и репрезентативность содержания теста проверяемой предметной области. Если в процессе конструирования теста принимали участие специалисты по данному предмету, следует указать их количество и профессиональную квалификацию. Если они выступали в роли экспертов при классификации заданий, необходимо привести дававшиеся им указания и коэффициент согласованности их мнений. Поскольку программы и содержание курсов со временем меняются, особенно желательно указать дату обращения к экспертам. Следует также сообщить число и характер проанализированных при подготовке теста программ и учебников, с указанием года издания.

Содержательная валидизация тестов учебных достижений обычно дополняется рядом эмпирических методов. И суммарный показатель, и выполнение отдельных заданий можно скорректировать относительно шкалы успеваемости. В общем, сохраняются те задания теста, которые показывают наибольший прирост процента учащихся, переходящих с более низких на более высокие уровни успеваемости. Другие дополнительные методы, когда они уместны, включают анализ типичных ошибок при выполнении учащимися теста и наблюдение за способами их работы. В последнем случае тестирование ведется в индивидуальном порядке, причем ученика просят при решении каждой задачи «рассуждать вслух». Существенность скоростного фактора может контролироваться по количеству тестируемых, не успевающих закончить тест, или с помощью одного из более тонких методов, обсуждавшихся в главе 4. Чтобы обнаружить возможное нежелательное влияние способности понять инструкцию на выполнение теста, можно вычислить коэффициент корреляции между показателями по данному тесту и показателями теста на понимание прочитанного. С другой стороны, если тест предназначен для оценки понимания текста, вопросы, относящиеся к содержанию еще не прочитанного отрывка, покажут, насколько испытуемый в состоянии на них ответить, исходя лишь из имеющихся у него предварительных знаний или пользуясь другими нерелевантными источниками информации (Scherich, & Hanna, 1977).

Области применения методов содержательной валидизации. Содержательная валидизация, особенно если она подкреплена такими эмпирическими проверками, как обсуждавшиеся выше, служит адекватным средством оценивания тестов достижений. Она позволяет ответить на два основных вопроса, касающихся валидности тес-

[1] Иначе говоря, *техническое задание на разработку теста. — Примеч. науч. ред.*

тов учебных и профессиональных достижений: 1) охватывает ли тест репрезентативную выборку конкретных навыков и знаний и 2) свободно ли выполнение теста от влияния посторонних факторов? Валидизация по содержанию особенно подходит для предметно-ориентированных тестов, описанных в главе 3. Поскольку выполнение таких тестов интерпретируется с точки зрения содержания предметной области или деятельности, совершенно очевидно, что содержательная валидизация является первейшим условием их эффективного использования. Однако и данные о других типах валидности не будут лишними, если мы хотим получить полную оценку эффективности таких тестов (см. Hambleton, 1984b).

Содержательная валидизация применима и к некоторым тестам, предназначенным для отбора и распределения профессиональных кадров, рассматриваемым в главе 17. Этот тип валидизации подходит в тех случаях, когда тест представляет собой выборочную проверку реальных рабочих операций или как-то иначе требует применения таких профессиональных навыков и знаний. В подобных случаях для доказательства близкого сходства между профессиональной деятельностью и тестом должен проводиться полный анализ содержания работы. Ясное, последовательное изложение применения этих методов валидизации в ходе разработки теста чтения для промышленности дано в одной из классических статей в этой области исследований (Schoenfeldt, Schoenfeldt, Acker, & Perlson, 1976). Работая в тесном контакте с занимающими разные должности лицами и их непосредственными начальниками, исследователи подробно изучили требования к чтению на нижних ступеньках служебной лестницы крупной промышленной компании с точки зрения содержания и уровня понимания. И только затем составлялись задания теста, которые полностью отвечали этим требованиям. Такой подход широко используется при разработке тестов для отбора государственных служащих как на федеральном уровне, так и на уровне штата (Hardt, Eyde, Primoff, & Tordy, 1981; Menne, McCarthy, & Menne, 1976; Primoff, & Eyde, 1988; Tordy, Eyde, Primoff, & Hardt, 1976).

С другой стороны, для тестов способностей и личности содержательная валидизация обычно не подходит и может даже уводить в сторону от правильного пути. Хотя рассмотрение релевантности и репрезентативности содержания должно быть составной частью начальных этапов конструирования любого теста, окончательная валидизация тестов способностей и личности требует эмпирической верификации с помощью методов, описанных в последующих разделах. Эти тесты не имеют того внутреннего сходства с выборочно оцениваемыми ими областями поведения, какое присуще тестам достижений. Следовательно, анализ их содержания может разве что выявить гипотезы, приведшие составителя к выбору определенного типа содержания для измерения заданного свойства. Такие гипотезы нужно еще эмпирически подтвердить, чтобы установить валидность оцениваемого теста.

В отличие от тестов достижений тесты способностей и личности не опираются на конкретный курс обучения или на общность предшествующего жизненного опыта, исходя из которых отбирается содержание теста достижений. Отсюда, способы выполнения разными людьми одних и тех же заданий в тестах способностей и личности, равно как и используемые ими при этом психологические процессы, могут существенно отличаться друг от друга. Таким образом, вполне возможно, что тот же самый тест у разных людей будет измерять различные функции, а это значит, что проверкой содержания теста фактически невозможно установить, какие психологические функции им измеряются. Так, выпускники колледжа могут решить некоторую задачу,

используя словесные формулировки или математические формулы, тогда как механик, возможно, придет к тому же решению путем пространственной визуализации. Или, например, тест, измеряющий способность к арифметическим рассуждениям у тех, кто только что перешел в среднюю школу, при предъявлении его студентам колледжа скорее всего выявит лишь индивидуальные различия в скорости вычислений.

Очевидная валидность. Содержательную валидность не следует смешивать с очевидной валидностью (*face validity*). Последняя, собственно, и не является валидностью в терминологическом смысле, ибо относится не к тому, что тест на самом деле измеряет, а к тому, что он при первом рассмотрении якобы измеряет. Очевидная валидность имеет отношение к тому, насколько тест «выглядит обоснованным» (т. е. валидным) для тех, кто его проходит, для тех, кто принимает ответственное решение о его использовании, да и вообще для всех неспециалистов. По существу, вопрос очевидной валидности касается «раппорта» и «паблик рилейшнз», т. е. налаживания взаимоотношений с тестируемыми и с общественностью. Хотя обычное употребление термина «валидность» в данной связи может вводить в заблуждение, сама по себе очевидная валидность — желательное свойство тестов. Например, когда тесты, первоначально предназначавшиеся для детей и разрабатывавшиеся применительно к школьной обстановке, вследствие их расширения впервые проводили на взрослых, те часто относились к таким тестам враждебно и критично именно из-за отсутствия очевидной валидности. В самом деле, если содержание теста представляется странным, неуместным, глупым или детским, результатом будет ухудшение сотрудничества, независимо от фактической валидности теста. Одной только объективной валидности теста явно недостаточно, особенно при тестировании взрослых. Очевидная валидность нужна тестам и для того, чтобы они эффективно функционировали в практических ситуациях. Она также влияет на степень приемлемости теста при вынесении законодательных и судебных решений, равно как и на оценку тестов широкой общественностью.

В новаторскую систематическую программу исследования тестирования как оно видится тестируемому (упоминавшуюся в главе 1) Барух Нево и его коллеги включили и изучение очевидной валидности (B. Nevo, 1985, 1992; B. Nevo, & Sfez, 1985). Сначала они привлекли внимание ученых к малому количеству исследований очевидной валидности, несмотря на ее возможный вклад в господствующее отношение к тестам. Затем они предложили количественную оценку очевидной валидности на основе оценок пригодности теста для его подразумеваемого применения, полученных от проходящих тестирование и других заинтересованных (но неискушенных в психометрике) лиц. Разработанные ими методы можно также использовать при оценивании отдельных заданий теста или, напротив, полных тестовых батарей. Опубликованные этими исследователями иллюстративные данные основывались на анализе ответов на Опросник обратной связи с экзаменуемым (*Examinee Feedback Questionnaire*), заполненный 1385 израильскими студентами, сдававшими вступительные экзамены в университет в форме шести письменных тестов. Результаты показали многообещающую согласованность ответов экзаменуемых, хорошую ретестовую надежность и дифференциацию тестов и подгрупп респондентов, планирующих специализацию в разных областях. Было рекомендовано регулярно сообщать в руководствах к тестам качественные и количественные данные, касающиеся очевидной валидности.

Очевидную валидность часто удается повысить простой переформулировкой заданий теста так, чтобы они выглядели уместными и правдоподобными в той конкретной обстановке, где предполагается использовать тест. Например, если тест, состоящий из простых арифметических задач, предназначен для квалифицированных рабочих механического цеха, то в условиях задач должны фигурировать машины или станки, а не количество апельсинов, которое можно купить на 86 центов, или иные предметы и персонажи из школьного задачника. Точно так же задания арифметического теста для военно-морского персонала можно сформулировать в морских терминах, не внося никакого изменения в измеряемые функции. Разумеется, очевидную валидность ни в коем случае нельзя считать заменой объективно устанавливаемой валидности. Нельзя рассчитывать на то, что улучшение очевидной валидности теста повысит его объективную валидность. Вместе с тем неправильно думать, что видоизменение теста, повышающее его очевидную валидность, никак не сказывается на его объективной валидности. Поэтому валидность теста в его окончательной форме всегда необходимо проверить заново, причем прямыми методами.

Методы предсказания критерия

Текущая и прогнозирующая валидизация. Методы установления валидности через предсказание критерия показывают эффективность теста в том, что касается прогнозирования выполнения индивидуумом точно определенной деятельности. Измерение критерия, относительно которого устанавливается валидность тестовых показателей, может производиться почти одновременно с ними или же через установленный промежуток времени. В зависимости от временных отношений между критерием и тестом *Стандарты тестирования* (1985) различают текущую и прогностическую валидности. Термин «прогнозирование» может использоваться как в широком смысле, означая предсказание по данному тесту в отношении любой критериальной ситуации, так и в более узком смысле предсказания в пределах некоторого временного интервала. В последнем смысле он и используется в выражении «прогностическая валидность». Информация, получаемая при прогнозирующей валидизации, особенно важна для тестов, используемых при отборе и распределении персонала. Прием на работу, отбор учащихся в колледжи или профессиональные училища, направление военнослужащих на курсы специальной подготовки — вот примеры ситуаций, требующих для принятия решений сведений о прогностической валидности используемых тестов. Сюда же можно отнести применение тестов в профотборе для отсеивания лиц, склонных в стрессовых ситуациях к эмоциональным расстройствам, и в психиатрической клинике — для назначения курса лечения, наиболее подходящего тем или иным пациентам.

В ряде случаев текущая валидность используется просто как заместитель прогностической валидности. На практике, для проведения прогнозирующей валидизации часто не хватает времени или не удается сформировать предварительную выборку, соответствующую целям тестирования. Поэтому в качестве компромиссного решения тесты проводятся на группе, для которой уже имеются данные по критерию. Например, тестовые показатели студентов колледжа могут сравниваться с их средней успеваемостью за период до момента тестирования, а тестовые показатели служащих — с их текущими производственными успехами.

Вместе с тем в определенных областях применения психологических тестов текущая валидность в наибольшей степени отвечает существу решаемых задач. Логическое различие между текущей и прогнозирующей валидизацией основано не на времени, а на целях тестирования. Текущая валидизация в полной мере применима к тестам, используемым для *диагноза* существующего положения дел, а не для предсказания будущих результатов. Это различие можно проиллюстрировать, задав два вопроса: «Является ли Смит достаточно квалифицированным летчиком?» и «Есть ли у Смита предпосылки к тому, чтобы стать квалифицированным летчиком?» Первый вопрос требует текущей валидизации соответствующего теста, второй — прогнозирующей валидизации.

Поскольку критерий для текущей валидизации всегда доступен во время тестирования, позволительно спросить, какую функцию в подобных ситуациях выполняет сам тест? В основном, такие тесты являются более простым, быстрым и дешевым заменителем критериальных данных. Например, если сбор данных о критерии требует постоянного наблюдения больного в стационаре в течение двух недель, то тест, позволяющий отделить норму от патологии и сомнительных случаев, мог бы заметно сократить число людей, занятых диагностическим наблюдением.

Ухудшение критерия. При определении валидности теста необходимо соблюдать меры предосторожности, с тем чтобы результаты теста не сказывались на положении тестируемого относительно выбранного критерия. Например, если преподавателю колледжа или мастеру на заводе станет известно, что данный студент или рабочий плохо справился с соответствующим тестом способностей, то это может плохо сказаться на оценке их деятельности. И наоборот, слишком высокие результаты по тесту могли бы подтолкнуть преподавателя или начальника к искусственному завышению академических оценок студентов или разряда рабочих соответственно. Такие влияния, очевидно, повышают корреляцию между показателями теста и критерием, которая, увы, не отражает действительного положения вещей.

Этот возможный источник ошибки при валидизации теста называют ухудшением или порчей критерия, поскольку оценки критерия «портятся» осведомленностью оценщика о тестовых показателях. Чтобы предотвратить действие такой ошибки, совершенно необходимо, чтобы лицам, производящим оценку критерия, ничего не было известно о тестовых результатах испытуемого. По этой причине тестовые показатели, используемые при «тестировании теста», должны держаться в строгом секрете. Порой трудно убедить преподавателей, работодателей, военное начальство и других официальных лиц в необходимости такой меры предосторожности. Стремясь использовать всю доступную информацию для принятия практических решений, эти люди могут не понимать того, что показателями теста нельзя пользоваться до тех пор, пока не будут получены критериальные данные и не будет проверена его валидность.

Меры критерия валидизации. Множество критериев, относительно которых может проводиться валидизация теста, соответствует множеству конкретных целей и областей его применения. Любой метод оценки поведения в любой ситуации мог бы дать критериальную меру для какой-то определенной цели тестирования. Однако критерии, относительно которых определяется приводимая в руководствах валидность тестов, можно разбить на несколько общих категорий. Для валидизации тестов интеллекта чаще всего используются тот или иной показатель *учебных достижений*

(academic achievement). Вот почему такие тесты иногда более точно характеризовали как средства измерения способности к обучению. В качестве конкретных показателей, используемых в роли меры критерия, выступают школьные оценки, показатели тестов достижений, сведения о переводе в следующий класс и об окончании школы, особые отличия и поощрения, а также интеллектуальные рейтинги учащихся, составляемые педагогами. Поскольку на эти рейтинги в значительной степени влияет результативность учебной деятельности каждого учащегося, постольку их, вероятно, можно отнести к категории мер критерия учебных достижений.

Различные показатели академических успехов использовались в качестве критериальных данных на всех уровнях обучения — от младших классов школы до колледжа и аспирантуры. Хотя их использовали главным образом для валидизации тестов общего интеллекта, они также служили критериями для некоторых тестов личности и комплексных батарей способностей. Например, при валидизации разнообразных тестов, предназначенных для отбора абитуриентов, общим критерием являлся средневзвешенный балл первокурсника. Эта мера представляет собой среднее из оценок по всем курсам первого года обучения, каждая из которых получает весовой коэффициент, соответствующий числу экзаменационных вопросов по курсу, за который она была получена.

Часто используемой разновидностью критерия академических достижений для неучащихся взрослых является объем полученного ими образования. Предполагается, что, в общем, люди с более высоким интеллектом продолжают свое образование, а менее интеллектуальные прекращают его раньше. Соображение, положенное в основу этого критерия, заключается в том, что образовательная лестница служит инструментом отбора с прогрессивно повышающимися требованиями, отсеивая на каждой ступени неспособных продолжать обучение. Хотя не подлежит сомнению, что, скажем, выпускники колледжа составляют группу, отобранную в соответствии с более высокими образовательными требованиями, чем окончившие начальную школу, связь между объемом образования и способностью к обучению весьма далека от полной. Экономические, социальные, мотивационные и другие неинтеллектуальные факторы могут влиять на продолжение человеком своего образования, особенно высшего. Кроме того, при такой текущей валидизации трудно решить, что является причиной, а что следствием. В какой степени полученные различия в показателях теста интеллекта есть просто результат разницы в образовании? И насколько точно тест мог бы предсказать индивидуальные различия в успехах при дальнейшем обучении? На эти вопросы можно ответить только в том случае, когда тест проводится до получения критериальных данных, как при прогнозирующей валидизации.

При разработке тестов специальных способностей в основу критерия валидизации часто кладут *эффективность специальной подготовки (performance in specialized training).* Например, валидность тестов механических способностей может устанавливаться относительно конечных результатов производственного обучения. Различные курсы бизнес-школ (машинописи, бухгалтерского учета и т. д.) обеспечивают критерии для тестов способностей в этих областях деятельности. Аналогично этому, результаты обучения в музыкальных или художественных училищах всегда использовались при валидизации тестов музыкальных и изобразительных способностей. Для ряда тестов профессиональных способностей валидизация проводилась относительно успешности обучения на юридическом, терапевтическом, стоматологическом и других факультетах университета. В случае изготавливаемых по особому заказу тестов, пред-

назначенных для использования в узкоспециальной программе тестирования, личные дела слушателей и курсантов часто служат источником критериальных данных. Яркий пример — валидизация тестов для отбора курсантов военных летных училищ относительно результатов начальной летной подготовки. Успешность выполнения программы специального обучения обычно используется и при валидизации других тестов, предназначенных для отбора военных и промышленных специалистов.

Среди показателей выполнения программы обучения, используемых в качестве критерия, можно упомянуть показатели тестов достижений, проводимых по завершении курсов, официально присваиваемые разряды и звания, оценки инструкторов и успешное окончание курсов в противоположность отчислению с них. Валидность комплексных батарей способностей часто устанавливалась относительно оценок по специальным предметам, проходимым в школе или в колледже. Например, показатели по тесту вербального понимания могут сравниваться с оценками по курсам родного языка, показатели по тесту пространственных представлений — с оценками по геометрии, и т. д.

В связи с использованием данных профессионального обучения в качестве мер критерия, полезно различать промежуточные и конечные критерии. При разработке теста для отбора курсантов военных летных училищ или теста медицинских способностей, например, конечными критериями были бы выполнение боевых заданий летчиком и достижение положительных результатов практикующим врачом соответственно. Очевидно, для получения таких критериальных данных потребовалось бы много времени. Сомнительно к тому же, что в реальной деятельности вообще можно получить действительно конечный критерий. Даже если бы такой конечный критерий в итоге оказался в нашем распоряжении, он, вероятно, подвергался действию множества неконтролируемых факторов, что сделало бы его относительно бесполезным. Например, было бы трудно оценить относительную степень успеха врачей различных специальностей, имеющих практику в разных частях страны. По этим причинам в качестве критериальных мер часто используются такие промежуточные критерии, как данные о результативности обучения на той или иной стадии.

Наилучшие во многих отношениях меры критерия валидизации основаны на последующем *выполнении реальной деятельности (job performance)*. В какой-то мере этот критерий использовался при валидизации тестов общего интеллекта и личности, но в значительно большей степени — при валидизации тестов специальных способностей. Кроме того, он обычно применяется для валидизации изготавливаемых по особому заказу тестов, касающихся отбора кадров для профессий, входящих в специальный перечень (авиадиспетчеры, операторы АЭС, инкассаторы и т. д.). Большинство мер выполнения профессиональной деятельности, не являясь, вероятно, конечными критериями, обеспечивают по крайней мере надежные промежуточные критерии для многих целей тестирования. В этом отношении они предпочтительнее данных о прохождении специального обучения. Вместе с тем при измерении выполнения той или иной работы не удается в такой степени стандартизовать условия, как в случае профессионального обучения. Более того, поскольку в этом случае требуется более длительный контроль за работающими, использование критерия выполнения реальной деятельности, вероятно, влечет за собой сокращение выборки валидизации. Ввиду того, что работники, занимающие номинально одинаковые должности, в разных организациях выполняют фактически неодинаковые функции, в руководстве к тесту вместе с данными о валидности относительно критерия реальной деятельности следует указать не

только использованные при валидизации конкретные меры этого критерия, но и дать краткую характеристику обязанностей, выполнявшихся этими работниками.

Валидизация методом *контрастных групп (contrasted groups)* обычно требует композиционного критерия, который отражает накапливающиеся и неконтролируемые селективные влияния повседневной жизни. Этот критерий, в конечном счете, основан на сохранении принадлежности индивидуума к конкретной группе в противоположность выбыванию из нее. Например, валидность теста музыкальных или механических способностей может проверяться сравнением показателей учащихся, зачисленных соответственно в музыкальную школу или на инженерно-механический факультет университета, с показателями тех, кто не выдержал требований этих учебных заведений. Разумеется, контрастные группы могут комплектоваться по любому критерию, такому как школьные оценки, рейтинги или выполнение нормы выработки, путем простого выбора крайних участков распределения соответствующих критериальных мер. Однако включаемые в данную категорию контрастные группы — это особые группы, которые становятся различными постепенно, под действием многочисленных требований повседневной жизни. В этом случае критерий оказывается более комплексным и менее поддающимся определению, чем ранее рассмотренные.

Метод контрастных групп довольно часто применяется при валидизации тестов личности. Так, при установлении валидности теста социальных качеств, можно было бы сравнить результаты тестирования торговых и административных работников, с одной стороны, с результатами тестирования конторских служащих и инженеров — с другой. Такое сравнение основывается на предположении, что те, кто выбрал профессии в сфере торговли или управления и продолжает там работать, отличаются как группа по своим социальным качествам от тех, кто предпочитает конторскую работу или инженерное дело. Аналогично, можно было бы сравнить тех студентов колледжа, кто принимал активное участие во внепрограммных мероприятиях, с теми, кто в течение сопоставимого периода пребывания в колледже ни разу в них не участвовал. Группы представителей различных профессий часто использовались при разработке и валидизации тестов интересов, таких как Бланк профессиональных интересов Стронга (*SVIB*), а также при подготовке шкал аттитюдов. Для определения валидности шкал аттитюдов иногда использовались группы, сформированные по политическому, религиозному, географическому и иным признакам, в отношении которых твердо известно, что они отражают противоположные точки зрения по определенным вопросам.

При эмпирической валидизации предметно-ориентированных тестов, в добавление к обычным методам валидизации по содержанию использовалось несколько адаптаций метода контрастных групп (Hambleton, 1984b). С этой целью группы, различающиеся по объему соответствующего обучения, сравнивались по результатам выполнения теста. При дихотомической оценке владения предметом проводился анализ четырехклеточных таблиц, в котором доля «зачетных» (*pass*) и «незачетных» (*fail*) показателей в необученной группе сравнивается с долей таких показателей в обученной группе (Panell, & Laabs, 1979). Аналогичные сравнения могут делаться и в тех случаях, когда тест предъявляется школьникам классом младше и классом старше того класса, в котором проходят конкретное понятие или формируется конкретное умение, оцениваемое данным тестом. Если доступны показатели за несколько разных периодов обучения, можно вычислить корреляцию между фактическим выполнением и объемом обучения.

При разработке некоторых тестов личности *психиатрический диагноз (psychiatric diagnosis)* используется и в качестве основания отбора заданий, и в качестве доказательства валидности теста. Такой диагноз может служить удовлетворительным критерием при условии, что он основан на длительном наблюдении и полной истории болезни, а не на беглом собеседовании или осмотре. В последнем случае на психиатрический диагноз можно положиться не больше чем на результат самого теста, и такой диагноз следует рассматривать не как критериальную меру, а как показатель или предсказатель, валидность которого еще должна быть установлена.

В связи с другими категориями критерия уже упоминались *рейтинги, или субъективные оценки (ratings),* даваемые школьными учителями, инструкторами специализированных курсов, мастерами на производстве. К ним можно добавить отзывы офицеров о действии подчиненных в штатных ситуациях, оценки учеников со стороны школьной администрации, оценки товарищей по работе, по классу, по клубу и т. д. Обсуждавшиеся до сих пор субъективные оценки представлялись лишь как вспомогательное средство получения информации о таких критериях, как академические достижения, эффективность специальной подготовки или успехи в работе. Теперь мы обращаемся к использованию субъективных оценок в качестве ядра критериальной меры. При таких условиях именно они задают значение критерия. Более того, такие оценки не ограничиваются описанием конкретных достижений, но включают личное суждение наблюдателя в отношении любого из множества свойств, на измерение которых ориентирован тест. Так, участников выборки валидизации наблюдатели могут ранжировать по таким признакам, как доминантность, искусность, оригинальность, лидерство или честность.

Подобные оценки использовались при валидизации почти всех типов тестов. Они особенно полезны в плане обеспечения критериев для тестов личности, поскольку установление объективных критериев в этой области связано с огромными трудностями. Это справедливо в отношении социальных качеств, так как их оценка основывается на личных контактах и потому может служить наиболее логически обоснованным критерием. Хотя эти оценки не свободны от ошибок, свойственных всем субъективным суждениям, они представляют собой ценный источник критериальных данных при условии их получения в тщательно контролируемых условиях. Способы повышения точности субъективных оценок и сокращения общих типов ошибок будут рассмотрены в главе 16.

Наконец, корреляции между новым тестом и *ранее доступными тестами (previously available tests)* часто приводятся в качестве доказательства валидности. Если новый тест представляет собой сокращенный или упрощенный вариант уже существующего теста, то последний можно с полным основанием считать критериальной мерой. Так, валидизация бланкового теста (типа «бумага—карандаш») может быть осуществлена относительно более сложно организованного и отнимающего много времени теста действия, валидность которого уже установлена. Или, скажем, валидность группового теста может устанавливаться относительно индивидуального теста. Тесты Стэнфорд—Бине, например, не раз служили критерием при валидизации групповых тестов. В таких ситуациях новый тест можно считать в лучшем случае грубой аппроксимацией ранее существующего. Следует отметить, что если новый тест не является более простым или более коротким заменителем ранее доступного теста, то использование последнего в качестве критерия недопустимо.

Существенное совершенствование конструирования тестов в 1980-е и 1990-е гг. привлекло внимание к *анализу критерия (criterion analysis)*. Это именно тот аспект работы по созданию теста, которым обычно пренебрегали в традиционных исследованиях тестов. На протяжении многих лет раздавались отдельные голоса, убеждавшие в необходимости систематических исследований критериев валидизации, однако практическое воплощение этих призывов было весьма скудным (L. R. James, 1973; Tenopyr, 1986). Даже в хорошо спланированных проектах, предполагавших тщательный анализ конкретного вида трудовой деятельности с целью получения ориентиров для разработки теста, результаты этого анализа практически не оказывали влияния на выбор меры критерия, используемого при последующей валидизации созданных вариантов теста. Обычно в качестве критерия принималось «то, что есть», и потому он часто был представлен одним общим показателем эффективности работы участников выборки валидизации, основанном на субъективных оценках начальства или на документах учета выработки.

В настоящее время широко признается, что валидность теста может быть наиболее эффективно исследована путем идентификации основных конструктов в выполнении определенной работы и последующего подбора или разработки тестов, показатели которых оценивают эти необходимые конструкты (J. P. Campbell, 1990 a; J. P. Campbell, McHenry, & Wise, 1990; L. V. Jones, & Applebaum, 1989; Messick, 1995). Замечательный пример применения всестороннего исследования критерия в качестве первого этапа разработки тестовой батареи дает Проект отбора и распределения специалистов сухопутных войск США (*U. S. Army's Selection and Classification Project*), больше известный под названием «Проект *А*» (J. P. Campbell, 1990b). Вследствие его общей значимости для применения тестов в сфере производства и управления этот крупномасштабный, семилетний проект более подробно рассматривается в главе 17.

Обобщение валидности. Прогностическая критериальная валидность (*criterion-prediction validity*) часто используется в локальных исследованиях валидизации, целью которых является оценка эффективности теста для какой-то конкретной программы. Этого подхода придерживаются в тех случаях, когда, например, некая компания хочет оценить тест для отбора кандидатов на одно из своих рабочих мест или когда некий колледж хочет выяснить, насколько хорошо тест академических способностей может предсказывать освоение определенного учебного курса его студентами. Прогностическую критериальную валидность можно лучше всего охарактеризовать как практическую валидность теста для строго определенной цели.

Когда в исследованиях валидизации на выборках работников промышленности показатели стандартизованных тестов способностей впервые попытались скоррелировать с результатами выполнения предположительно родственных видов работы, была обнаружена значительная вариация коэффициентов валидности (Ghiselli, 1959, 1966). Аналогичная вариабельность коэффициентов валидности наблюдалась и тогда, когда критериями служили оценки по различным учебным предметам (G. K. Bennett, Seashore, & Wesman, 1984). Такие результаты привели к общему пессимизму в отношении обобщимости валидности теста на различные ситуации. До середины 1970-х гг. «ситуационная специфичность» психологических требований обычно считалась серьезным ограничением применимости стандартизованных тестов в профотборе. Однако Шмидт, Хантер и их коллеги с помощью тонкого статистического анализа этой проблемы показали, что большая часть дисперсии полученных коэффициентов

валидности может быть просто статистическим артефактом, возникающим вследствие малого объема выборки, ненадежности критерия и ограничения диапазона изменчивости в выборках работников.[1]

Выборки работников предприятий, доступные исследователям при валидизации тестов, обычно слишком малы, чтобы дать устойчивую оценку корреляции между прогнозирующим показателем и критерием. По той же причине получаемые коэффициенты могут оказаться слишком низкими, чтобы достичь статистической значимости в используемой для валидизации выборке, и потому не пригодными в качестве доказательства валидности теста. По имеющимся оценкам примерно половина выборок работников промышленных предприятий, используемых в исследованиях валидности, включает не более 40–50 человек (Schmidt, Hunter, & Urry, 1976). При таких малых выборках валидизации через предсказание критерия технически не осуществима.

Применяя свои недавно разработанные методы анализа к данным многих выборок, извлеченных из большой совокупности работников промышленности, Шмидт, Хантер и их сотрудники сумели показать, что валидность тестов вербальных, числовых и логических способностей можно распространить на значительно более широкий круг профессий, чем считалось ранее. Было доказано, что дисперсия коэффициентов валидности, обычно обнаруживавшаяся в более ранних исследованиях валидизации на выборках работников промышленности, не превышала величины случайной изменчивости. Этот вывод остается справедливым, даже когда специфические функции работников, казалось бы, существенно различаются в зависимости от места и характера работы. В конечном счете, успешное выполнение самых разных профессиональных задач во многом зависит от общего ядра когнитивных умений. Включенные в эти исследования тесты охватывали, главным образом, содержание и умения того типа, которые выборочно проверяются традиционными тестами интеллекта и академических способностей. Может показаться, что этот кластер когнитивных умений и знаний должен обладать значительной прогнозирующей силой в отношении выполнения разнообразной учебной и профессиональной деятельности, спрос на которую существует в обществах с передовой технологией. Однако более точных решений при отборе персонала обычно удается достичь при рассмотрении показателей по двум-трем широким когнитивным кластерам, предпочтительно дополненных замерами трудовых навыков предназначенных для выполнения конкретных профессиональных задач (Hartigan, & Wigdor, 1989; L. L. Wise, McHenry, & Campbell, 1990; Zeidner, & Johnson, 1991).

Метаанализ. Статистические методы, используемые при изучении пределов обобщимости валидности, по существу дают нам способ объединения данных из различных исследований. С их помощью можно объединять данные прошлых и настоящих исследований, проведенных в одном или в разных местах, а также привлекать информацию из доступных публикаций. Хотя эта группа методов была внедрена в психологические исследования и впервые названа метаанализом (*meta-analysis*) в 1970-х гг.

---

[1] Эта работа была частью длительной программы исследований, результаты которых отражены во многих статьях и монографиях. К числу наиболее важных с точки зрения обсуждаемого здесь вопроса относятся следующие публикации: Pearlman, Schmidt, & Hunter (1980), Schmidt, Gast-Rosenberg, & Hunter (1980), Schmidt & Hunter (1977), Schmidt, Hunter, & Pearlman (1981), Schmidt, Hunter, Pearlman, & Shane (1979).

(Glass, 1976; Schmidt, & Hunter, 1977), лежащие в их основе вычислительные процедуры использовались уже в течение нескольких десятилетий, особенно в других науках (Hartigan, & Wigdor, 1989, chap. 6). Метаанализ получил растущее признание в психологии как возможная замена традиционных литературных обзоров (Lipsey, & Wilson, 1993; Schmidt, 1992). Такие обзоры, как правило, содержали информацию о тех исследованиях, в которых получены статистически значимые результаты, касающиеся, например, различий между средними контрольных и экспериментальных групп или корреляций между тестовыми показателями и другими переменными. При таком подходе многообещающие позитивные результаты часто терялись в силу того, что используемые в отдельных исследованиях выборки были слишком малы, чтобы обеспечить получение значимых различий.

Благодаря объединению опубликованных данных нескольких исследований и приписыванию им весов (насколько это возможно) на основе релевантных методологических и вещественных признаков каждого исследования, метаанализ может выявить важные позитивные результаты. Дополнительное преимущество метаанализа состоит в том, что он допускает вычисление *величины эффектов (effect sizes)*. И по теоретическим, и по практическим соображениям оценка величины различия или корреляции гораздо полезнее простой демонстрации их статистически значимого отличия от нуля.

Два последних десятилетия XX в. свидетельствовали о быстром росте числа метааналитических исследований почти во всех областях психологии. Приложения метаанализа в исследованиях проблем профотбора и распределения персонала, вероятно, привлекли самое широкое внимание (см. главу 17). Интерес к метаанализу неуклонно растет и, соответственно, постоянно совершенствуются его процедуры. Хотя некоторые приемы метаанализа считаются спорными, основные результаты, получаемые с помощью разных его процедур, практически не различаются.[1]

Методы идентификации конструкта

Термин «конструктная валидность» (*construct validity*) был официально введен в лексикон психометристов в 1954 г., ознаменованном выходом в свет *Технических рекомендаций для психологических тестов и диагностических методик (Technical Recommendations for Psychological Tests and Diagnostic Techniques*, — первого издания современных *Стандартов тестирования*. Первое подробное описание конструктной валидности появилось в следующем году в статье Кронбаха и Мила (Cronbach, & Meehl, 1955). Дискуссии вокруг понятия конструктной валидности, развернувшиеся сразу после этой публикации и ведущиеся с неослабной энергией до сих пор, способствовали прояснению исходных предпосылок, лежащих в основе методов установления этого типа валидности, и обеспечению систематического обоснования их использования.

[1] Современные приложения, подробное объяснение способов и критические оценки метаанализа можно найти в следующих работах: Hartigan & Wigdor (1989), Hedges (1988), Hunter & Schmidt (1990), L. R. James, Demaree, Mulaik, & Ladd (1992), L. V. Jones & Applebaum (1989), R. Rosenthal (1991), Schmidt (1992), Schmidt et al. (1993), Schmidt, Ones, & Hunter (1992). Что касается простого введения в статистические процедуры метаанализа, см. F. M. Wolf (1986). Более широкая перспектива использования метаанализа в поведенческих науках представлена в Cook et al. (1992), Cooper & Hedges (1994), Hasselblad & Hedges (1995), Wachter & Straf (1990).

Валидизация конструкта привлекла внимание к роли психологической теории в конструировании тестов и к необходимости формулировать гипотезы, которые можно было бы подтвердить или опровергнуть в процессе валидизации теста. Понятие конструктной валидности к тому же стимулировало поиск новых способов сбора данных о валидности. Хотя некоторые из этих способов были уже давно известны, их область применения была существенно расширена, чтобы иметь возможность включить большее число конкретных процедур.

Конструктная валидность теста показывает, насколько его результаты могут рассматриваться в качестве меры некоего теоретического конструкта или свойства. Примерами таких конструктов являются академические способности, понимание механических закономерностей, беглость речи, скорость ходьбы, нейротизм и тревожность. Каждый конструкт разрабатывается в целях объяснения и организации наблюдаемых последовательностей реакций. Он выводится из установленных взаимосвязей между поведенческими характеристиками. Валидизация конструкта требует постепенного накопления информации из разных источников. В дело идут любые данные, проливающие свет на природу рассматриваемого свойства и на условия, от которых зависит его развитие и проявление. Примеры конкретных методов, способствующих идентификации конструктов, рассматриваются ниже.

Возрастные изменения. Главным критерием, используемым при валидизации ряда традиционных тестов интеллекта, является *возрастная дифференциация (age differentiation).* Такие тесты, как шкала Стэнфорд—Бине и большинство тестов для дошкольников, проверяются на соответствие хронологическому возрасту, с тем чтобы выяснить, повышаются ли тестовые показатели детей от года к году. Поскольку ожидается, что способности и умения детей возрастают с каждым годом, предполагается, что и показатели теста должны соответственно повышаться, если этот тест является валидным. Само понятие возрастной шкалы интеллекта, введенное А. Бине, основано на допущении, что «интеллект» увеличивается с возрастом, по крайней мере до наступления зрелости.

Критерий возрастной дифференциации, разумеется, неприменим к таким функциям, которые не обнаруживают четких и последовательных возрастных изменений. В области измерения личности, например, этот критерий нашел ограниченное применение. Кроме того, следует отметить, что возрастная дифференциация, даже когда она применима, является необходимым, но не достаточным условием валидности. Так, если тестовые показатели не улучшаются с возрастом, такой результат, вероятно, указывает на то, что данный тест не является валидной мерой способностей, которые он должен выборочно проверять. С другой стороны, доказательство того, что тест измеряет нечто, увеличивающееся с возрастом, еще не дает достаточно точного определения области, охватываемой этим тестом. Замеры роста или веса будут также обнаруживать регулярные прибавки с возрастом, хотя и производятся отнюдь не тестом интеллекта.

В заключение подчеркнем еще один момент, касающийся интерпретации возрастного критерия. Психологический тест, валидность которого установлена относительно такого критерия, измеряет характерные черты поведения, усиливающиеся с возрастом в условиях той среды, в которой тест был стандартизован. Поскольку различные культуры могут стимулировать и поощрять развитие непохожих черт поведения, критерий возрастной дифференциации нельзя считать универсальным. Как и все другие критерии, он действителен лишь для определенной культурной среды.

Анализ возрастных изменений является также основным методом конструктной валидизации порядковых шкал Пиаже, обсуждаемых в главах 3 и 9. В основу таких шкал положено допущение о *последовательном структурировании (sequential patterning)* развития, согласно которому достижение более ранних стадий в развитии понятий служит необходимой предпосылкой к приобретению более поздних когнитивных умений. Таким образом, содержанию этих шкал присуща имманентная иерархичность. Конструктная валидизация порядковых шкал, следовательно, включает эмпирические данные о неизменности последовательных ступеней развития. Это предполагает проверку выполнения теста детьми на разных уровнях развития любого исследуемого понятия, например сохранения или постоянства объекта. Иначе говоря, необходимо установить, действительно ли дети, владеющие определенным понятием на данном уровне, владеют им и на более низких уровнях.

Корреляции с другими тестами. Корреляции между новым и аналогичными ему существующими тестами иногда рассматриваются как доказательство того, что новый тест измеряет примерно ту же сферу поведения, что и другие одноименные тесты, такие как тесты интеллекта или тесты механических способностей и т. д. В отличие от корреляций, получаемых при установлении прогностической критериальной валидности, эти корреляции должны быть умеренно высокими. Если новый тест слишком тесно коррелирует с уже существующим и не обладает такими дополнительными преимуществами, как краткость или легкость проведения, то это означает излишнее дублирование имеющегося теста.

Корреляции с другими тестами используются, помимо этого, в качестве меры относительной свободы нового теста от влияния определенных посторонних факторов. Например, тесты специальных способностей или личности не должны иметь высоких корреляций с тестами общего интеллекта или академических способностей. Точно так же понимание читаемого не должно заметно влиять на выполнение таких тестов. Это объясняет, почему корреляции с тестами общего интеллекта, чтения и вербального понимания иногда приводят в качестве косвенного, или негативного, доказательства валидности. В этих случаях высокие корреляции ставили бы под сомнение валидность теста. Однако низкая корреляция сама по себе еще не гарантирует достаточной валидности. Нужно иметь в виду, что это использование корреляций с другими тестами аналогично одному из рассмотренных выше вспомогательных приемов валидизации через описание содержания.

Факторный анализ. Разработанный как средство идентификации психологических черт, факторный анализ имеет самое прямое отношение к методам валидизации конструкта. В сущности, факторный анализ представляет собой тонкий статистический инструмент анализа взаимосвязей данных о поведении. Например, если 300 человек прошли 20 тестов, то первый шаг состоит в вычислении попарных корреляций между всеми тестами. Простой просмотр итоговой матрицы из 190 коэффициентов корреляции уже мог бы выявить некоторые группы (кластеры) коррелирующих между собой тестов, что означало бы обнаружение общих черт. Так, если такие тесты, как словарный, аналогий, антонимов и завершения предложений, тесно коррелируют между собой и слабо — со всеми другими тестами, то мы могли бы, в предварительном порядке, вывести наличие фактора вербального понимания. Поскольку анализ корреляционной матрицы визуальным путем и труден и ненадежен, то для обнаружения

общих факторов, необходимых для объяснения полученных корреляций, были разработаны более точные статистические методы. Эти методы факторного анализа будут еще рассмотрены в главах 11, в связи с их использованием в исследованиях природы интеллекта, где они и зародились.

В ходе факторного анализа равное количеству тестов число переменных или категорий, с помощью которых описываются результаты каждого тестируемого, сокращается до нескольких факторов или общих черт. В приведенном выше примере для объяснения попарных корреляций между 20 тестами могло бы хватить 5 или 6 факторов. Иначе говоря, описание каждого человека с помощью показателей по 20 тестам можно было бы заменить характеристикой на основе оценок по 5 или 6 факторам. Главное назначение факторного анализа состоит в упрощении описания поведения путем сокращения большого числа разнообразных категорий (соответствующих тестируемым переменным) до нескольких общих факторов, или черт.

После идентификации факторов их можно использовать для описания факторной структуры теста. Каждый тест можно, таким образом, охарактеризовать исходя из главных факторов, определяющих его показатели, с учетом веса или нагрузки каждого фактора и корреляции теста с каждым из них. Такую корреляцию иногда приводят как *факторную валидность (factorial validity)* теста. Так, если фактор вербального понимания имеет корреляцию 0,66 со словарным тестом, то факторная валидность этого теста как средства измерения вербального понимания равна 0,66. Следует отметить, что факторная валидность по существу представляет собой корреляцию теста со всем тем, что есть общего у группы тестов или других индексов поведения. Анализируемое множество переменных может, разумеется, включать в себя как данные тестов, так и данные иного рода. Субъективные оценки (*ratings*) и другие меры критерия, наряду с другими тестами, могут быть использованы для исследования факторной структуры конкретного теста и для определения измеряемых им общих черт.

Внутренняя согласованность. В публикуемой информации о некоторых тестах, особенно применяемых для исследования личности, можно встретить утверждение, что валидность теста была установлена методом внутренней согласованности. Существенной особенностью этого метода является использование в качестве критерия валидизации суммарного показателя самого теста. Иногда для оценки внутренней согласованности теста приспосабливается метод контрастных групп, которые в этом случае формируются из испытуемых с самыми высокими и с самыми низкими суммарными показателями по данному тесту. Результаты выполнения каждого задания теста группой с верхним значением критерия сравнивается затем с соответствующими результатами группы с нижним значением критерия. Задания, по которым не удалось обнаружить существенно большей доли «правильных» (совпадающих с ключом) ответов в группе с верхним значением критерия по сравнению с группой с низким значением критерия, признаются невалидными и либо отбрасываются, либо перерабатываются. Можно также воспользоваться корреляционными методами, например вычислить бисериальные коэффициенты корреляции между исходами («справился — не справился») каждого задания и суммарным показателем теста. В этом случае сохраняются только те задания, для которых отмечена значимая корреляция с тестом в целом. Если тест состоит из заданий, прошедших такого рода отбор, то можно говорить о его внутренней согласованности, поскольку каждое его задание дифференцирует респонденов в том же направлении, что и тест в целом.

Еще одно применение критерия внутренней согласованности связано с корреляцией между показателями субтестов и суммарным показателем теста. Многие тесты интеллекта, например, состоят из раздельно проводимых субтестов (таких, как словарный, арифметический, недостающие детали и т. д.), показатели которых складываются при нахождении суммарного тестового показателя. При конструировании этих тестов показатели по каждому субтесту часто коррелируются с суммарным показателем, и субтесты, имеющие низкую корреляцию с тестом в целом, исключаются. Коэффициенты корреляции оставшихся субтестов с суммарным показателем теста приводятся затем как свидетельство внутренней согласованности всего этого измерительного инструмента.

Очевидно, что корреляции, отражающие внутреннюю согласованность теста, являются по существу мерой его однородности. Поскольку это свойство помогает охарактеризовать область поведения или отдельную черту, выборочно проверяемые тестом, то степень однородности теста имеет отношение к его конструктной валидности. Тем не менее вклад данных о внутренней согласованности теста в его валидизацию носит ограниченный характер. При отсутствии внешних по отношению к тесту данных мало что можно узнать о том, что он в действительности измеряет.

Конвергентная и дискриминантная валидизация. В своем глубоком анализе методов валидизации конструкта Д. Т. Кэмпбелл (D. T. Campbell, 1960) обратил внимание на следующее: для доказательства конструктной валидности мы должны показать, что тест не только имеет высокие корреляции с другими переменными, с которыми он должен коррелировать исходя из теоретических предположений, но и *не* имеет значимых корреляций с переменными, от которых он должен отличаться. В своей более ранней статье Кэмпбелл и Фиске (D. T. Campbell, & Fiske, 1959) охарактеризовали первый и второй аспект анализа конструктной валидности как конвергентную и дискриминантную валидизацию соответственно. Корреляция показателей теста количественных рассуждений с последующими оценками по курсу математики могла бы служить примером конвергентной валидизации. Для того же самого теста одним из доказательств его дискриминантной валидности могло бы быть получение низкой и статистически незначимой корреляции с тестом понимания текста, поскольку умение читать не является релевантной переменной для теста, предназначенного измерять количественные рассуждения.

Напомним, что требование низкой корреляции с нерелевантными тесту переменными рассматривалось выше в связи с дополнительными и превентивными мерами, рекомендуемыми при установления содержательной валидности. Кроме того, дискриминантная валидизация особенно важна при установлении валидности тестов личности, в которых нерелевантные переменные могут влиять на результаты самым непредсказуемым образом.

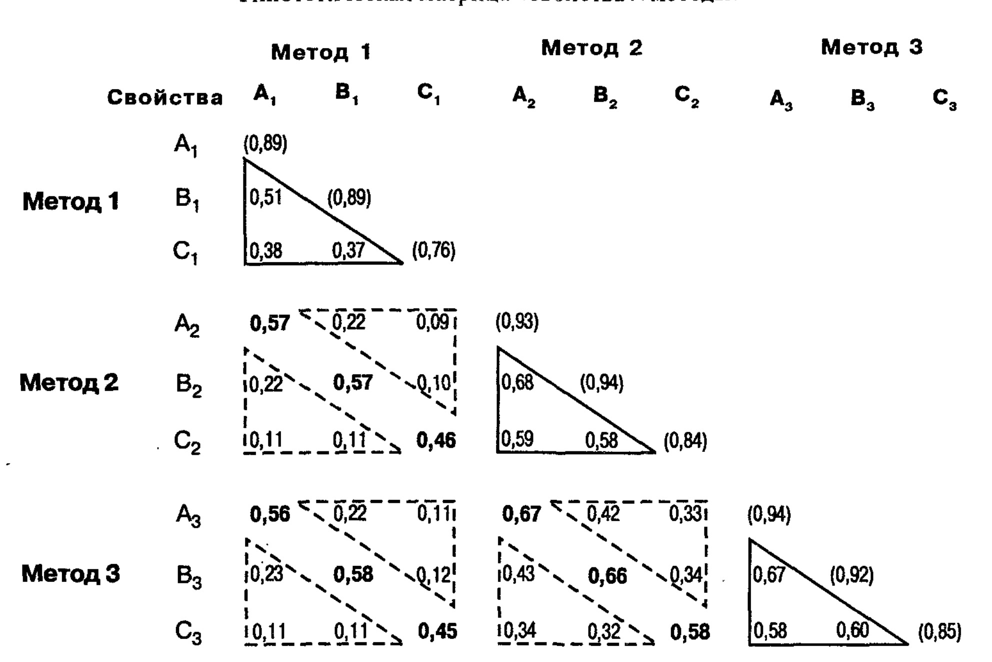
В упомянутой выше статье (D. T. Campbell, & Fiske, 1959) предложен систематизированный экспериментальный план для одновременного проведения конвергентной и дискриминантной валидизации тестов, названный авторами *матрицей «свойства × методы» (multitrait-multimethod matrix)*. По существу дела, этот план предполагает оценку двух или более свойств двумя или более методами. Гипотетический пример, взятый из этой статьи, поможет прояснить предлагаемый подход. В табл. 5–1 показаны все возможные корреляции между показателями, полученными при измерении каждого из трех свойств тремя методами. Эти свойства *A, B* и *C* могли бы быть, ска-

жем, тремя такими свойствами личности, как *A*) доминантность, *B*) общительность и *C*) мотивация достижения. В качестве методов могли бы использоваться: 1) опросник, заполняемый респондентом, 2) проективная методика и 3) оценки сверстников. При этих условиях A_1 служит обозначением показателей доминантности, полученных с помощью опросника, A_2 представляет показатели доминантности по проективному тесту, а C_3 — оценки мотивации достижения, даваемые сверстниками.

Гипотетические коэффициенты корреляции, приведенные в табл. 5–1, включают в себя коэффициенты надежности (они стоят в скобках вдоль главной диагонали) и коэффициенты валидности (напечатаны полужирным шрифтом вдоль трех более коротких диагоналей). Как показывают коэффициенты валидности, результаты измерения каждого свойства различными методами коррелируют между собой. Таким образом каждая мера проверяется на соответствие другим, независимым мерам того же свойства, как и в знакомой нам процедуре валидизации. Таблица также содержит коэффициенты корреляции между *разными* свойствами, измеренными *одним* (сплош-

Таблица 5–1

Гипотетическая матрица «свойства × методы»



	Свойства	Метод 1 A_1	Метод 1 B_1	Метод 1 C_1	Метод 2 A_2	Метод 2 B_2	Метод 2 C_2	Метод 3 A_3	Метод 3 B_3	Метод 3 C_3
Метод 1	A_1	(0,89)								
	B_1	0,51	(0,89)							
	C_1	0,38	0,37	(0,76)						
Метод 2	A_2	**0,57**	0,22	0,09	(0,93)					
	B_2	0,22	**0,57**	0,10	0,68	(0,94)				
	C_2	0,11	0,11	**0,46**	0,59	0,58	(0,84)			
Метод 3	A_3	**0,56**	0,22	0,11	**0,67**	0,42	0,33	(0,94)		
	B_3	0,23	**0,58**	0,12	0,43	**0,66**	0,34	0,67	(0,92)	
	C_3	0,11	0,11	**0,45**	0,34	0,32	**0,58**	0,58	0,60	(0,85)

Примечание. Буквами А, В и С обозначены свойства, а нижними индексами — методы. Коэффициенты валидности (корреляции между показателями одного свойства, измеренного разными методами) представлены тремя диагонально расположенными наборами чисел, напечатанных полужирным шрифтом. Коэффициенты надежности (корреляции между показателями одного свойства при его повторном измерении тем же методом) представлены числами в скобках вдоль главной диагонали. В треугольниках из сплошных линий заключены корреляции между разными свойствами, измеренными одним методом; в пунктирных треугольниках — корреляции между разными свойствами, измеренными разными методами.

(Из Campbell & Fiske, 1959, p. 82.
Copyright 1959 by the American Psychological Association. Воспроизведено с разрешения)

ные треугольники) методом, и *разными* свойствами, измеренными *разными* (пунктирные треугольники) методами. Конструктная валидность может считаться удовлетворительной, если коэффициенты валидности явно выше коэффициентов корреляции между разными свойствами, измеренными разными методами; они также должны быть выше коэффициентов корреляции между разными свойствами, измеренными одним методом. Например, корреляция между показателями доминантности по опроснику и по проективной методике должна быть выше корреляции между показателями доминантности и общительности по опроснику, заполняемому самим испытуемым. Если бы последняя корреляция, отражающая дисперсию общего метода, оказалась высокой, это могло бы означать, например, что на показатели респондента по этому опроснику чрезмерно влияет какой-то нерелевантный общий фактор, такой как способность понимать вопросы или желание представить себя в выгодном свете по всем свойствам.

Экспериментальные вмешательства. Еще один источник данных для валидизации конструкта обеспечивают эксперименты, в которых исследуется влияние выбранных переменных на показатели теста. При проверке валидности теста, предназначенного, например, для использования в программе индивидуализированного обучения, есть только один путь — сравнить показатели тестирования до и после экспериментального обучения. Логическое обоснование такого теста требует низких показателей при первом тестировании, проводимом до соответствующего обучения, и высоких показателей при втором тестировании, после обучения. То же соотношение может проверяться и для отдельных заданий теста. В идеале с каждым заданием до обучения должно справиться минимальное, а после обучения — максимальное число учеников. Задания, с которыми мало кто справляется в обоих случаях, слишком трудны, а те, с которыми справляются почти все и до и после обучения, слишком доступны с точки зрения целей, преследуемых тестом. Если же многие в первый раз справляются, а во второй раз не справляются с заданием, то что-то неладно или с этим заданием, или с обучением, или с тем и другим.

Тест, предназначенный для измерения склонности к тревоге (*anxiety-proneness*), можно проверить, давая его испытуемым до и после того, как они были помещены в обстановку, провоцирующую состояние тревоги (примером может служить проверка знаний в напряженных или мешающих выполнению задания условиях). Исходные тестовые показатели тревожности можно затем соотнести с физиологическими и иными показателями выражения тревоги во время и после экспериментального воздействия. Другую (дифференциальную) гипотезу в отношении теста тревожности можно оценить, проводя тест до и после вызывающего тревогу события и наблюдая за тем, происходит ли существенное увеличение тестовых показателей при втором тестировании. Положительные результаты такого эксперимента будет свидетельствовать о том, что тестовые показатели отражают текущий уровень тревожности. Аналогичным образом можно планировать эксперименты для проверки гипотез относительно любой конкретной черты, измеряемой данным тестом.

Моделирование структурными уравнениями. В добавление к идентификации конструктов и в тестовых показателях, и в критериальной деятельности, важным продвижением вперед в области валидизации тестов стало рассмотрение отношений между конструктами и того пути, по которому осуществляется влияние конструкта на

выбранную в качестве критерия деятельность (J. P. Campbell, 1990a; Messick, 1989; Schmidt, Hunter, & Outerbridge, 1986). Например, интерес человека к конкретной области может влиять на эффективность его работы через повышение усвоения релевантных фактуальных знаний, через приобретение требуемых процедурных навыков или через развитие мотивации, необходимой для того, чтобы проявлять максимум усилий и выдерживать напряжение при выполнении производственных заданий в установленные сроки. Выяснение того, *каким образом* идентифицированный конструкт или индивидуальная особенность приводит к хорошим или плохим результатам, вносит существенный вклад в понимание того, почему тест имеет высокую или низкую валидность в данной ситуации Такой анализ значительно облегчается при использовании статистического метода, называемого моделированием структурными уравнениями. Применение этого метода резко возросло в 1980-х и 1990-х гг., о чем свидетельствует, например, создание специального журнала — *Structural Equation Modeling* (1994). Данный метод тесно связан с различными версиями путевого анализа, а оба этих подхода часто называют (нестрого) «причинным моделированием».[1]

Каких конкретных результатов позволяет достичь моделирование структурными уравнениями и как оно возникло? При изучении элементарной статистики студенты быстро узнают, что корреляция не показывает причинной связи. Знакомый всем пример — фиктивная корреляция возраста. В смешанной выборке школьников в возрасте от 6 до 14 лет корреляция между ростом и умением производить арифметические вычисления скорее всего будет высокой, но мы вряд ли сделаем из этого вывод, что какая-то из этих переменных влияет на другую. Эта корреляция появляется, в основном, за счет изменения возраста, который, в свою очередь, связан с объемом полученного образования. Пытаясь разобраться в причинных связях, исследователи начали в 1960-х — 1970-х гг. использовать перекрестные с лагом планы эксперимента (*cross-lagged experimental design*) (D. T. Campbell, & Stanley, 1966; Cook, & Campbell, 1976, p. 284–293). Например, чтобы проанализировать причинные взаимосвязи между отношением ученика к математике и его показателями в этой области знаний, замеры отношения и достижений можно было бы произвести в два разных момента времени. Затем можно было бы вычислить перекрестную корреляцию между отношением к математике в момент t_1 и достижениями в математике в момент t_2 и между достижениями в математике в момент t_1 и отношением к математике в момент t_2. Величина этих двух коэффициентов корреляции, вероятно, должна показывать относительную силу влияния в обоих направлениях. В течение ряда лет этот план казался многообещающим способом оценки воздействия двух переменных друг на друга.

Вскоре, однако, в ходе логического и статистического анализов были обнаружены серьезные недостатки метода перекрестных с лагом корреляций. Хотя сама по себе схема перекрестных сравнений через заданный интервал времени, положенная в основу экспериментального плана, не может вызвать никаких упреков, использование простейших корреляций нулевого порядка, вероятно, искажает результаты эксперимента и ведет к некорректным выводам о причинных связях (Rogosa, 1980). Источники ошибок в этой процедуре связаны с неспособностью учесть, во-первых, корреляции

[1] Чтобы избежать философских выводов и допущений о первопричине или полной причинной цепи любого события, психологи предпочли более нейтральные выражения, наподобие того, что *A* определяет, влияет или воздействует на *B*. Тем не менее термин «причинный» иногда употребляют для ссылки на все эти связи и отношения, предполагая понимание его ограничений читателями (см., например, L. R. James, Mulaik, & Brett, 1982, chap. 1; P. A. White, 1990).

между начальными и конечными значениями переменных; во-вторых, надежность самих переменных и их временную устойчивость; и, в-третьих, возможное влияние неизмеряемых переменных, таких как возраст и объем полученного образования в упоминавшемся выше классическом примере. Моделирование структурными уравнениями свободно от подобных упреков. По существу, это достигается благодаря применению уравнений регрессии для предсказания значений зависимых переменных по независимым переменным в различных моделях причинного анализа, включая перекрестные измерения с лагом. В этом методе для нахождения коэффициентов регрессии используются частные (парциальные) корреляции, в результате чего в уравнение вводятся все связи между переменными; учитываются как ошибки измерения, так и ошибки выборки; наконец, принимаются некоторые меры предосторожности, с тем чтобы по крайней мере признать возможность влияния дополнительных, неизмеряемых причинных переменных (Bentler, 1988; L. R. James et al., 1982; Loehlin, 1992; Rogosa, 1979).

Первый этап моделирования структурными уравнениями — построение модели гипотетических причинных отношений, которую предстоит проверить. Важно, чтобы эта модель основывалась на доскональном знании существующей информации о переменных и изучаемой ситуации. Включаемые в модель гипотетические отношения должны иметь прочное теоретическое обоснование. Проверка модели осуществляется путем решения системы совместных линейных уравнений регрессии.[1] В причинном моделировании число уравнений обычно больше числа неизвестных, что позволяет получить решения для нескольких альтернативных моделей. Каждая модель сравнивается с исходной, эмпирической корреляционной матрицей для определения степени согласия. При этом, однако, несколько причинных моделей могут обнаружить примерно одинаковую степень согласия с эмпирическими данными (MacCallum, Wegener, Uchino, & Fabrigar, 1993). Такие статистически эквивалентные модели могут представлять различные причинные пути и, следовательно, давать альтернативные объяснения наблюдаемых эффектов. Опираясь на свое знание изучаемой ситуации, исследователь должен оценить эти альтернативные модели с точки зрения их правдоподобия и физического смысла.

Другая особенность моделирования структурными уравнениями состоит в том, что здесь оцениваются, как правило, причинные отношения между конструктами, а не между отдельными измеряемыми переменными. Например, для определения отношения учащегося к математике можно было бы использовать ряд показателей, таких как меры интереса, целеустремленности, представления о собственных математических способностях и других релевантных аффективных переменных. Тогда общая изменчивость этих показателей определяла бы конструкт отношения учащегося к математике, который можно связать с его последующими математическими достижениями. Использование конструктов обеспечивает более устойчивые и надежные оценки, в которых ошибка и специфические дисперсии отдельных показателей сводятся на нет.

---

[1] Для ознакомления с деталями этого метода см. Bollen (1989) и Loehlin (1992). Что касается реальных вычислений, то можно воспользоваться любой из имеющихся в наличии компьютерных программ, например *LISREL* (Hayduk, 1988; Jöreskog & Sörbom, 1986, 1989) и *EQS* (Bentler, 1985). [На рус. яз. см. соответственно: *Хейс Д.* Причинный анализ в статистических исследованиях: Пер. с англ. — М.: Финансы и статистика, 1981; *Боровиков В. П., Боровиков И. П.* STATISTICA® — Статистический анализ и обработка данных в среде Windows®. — М.: Филин, 1997. — С. 528–565. — *Примеч. науч. ред.*]

В настоящее время существует несколько методологических подходов к моделированию структурными уравнениями, так же как и целый ряд модификаций и процедурных усовершенствований этого метода (см., например, Anderson, & Gerbing, 1988; Bentler, 1990; Bollen, & Long, 1993; Breckler, 1990; Cole, Maxwell, Arvey, & Salas, 1993; James, 1980; Mulaik et al., 1989). И хотя моделирование структурными уравнениями все еще находится в стадии развития, этот метод является многообещающим в плане объединения теоретического, экспериментального и статистического подходов. Он уже нашел широкое применение для решения проблем психологии личности, возрастной, социальной, промышленной и педагогической психологии (например, Graves, & Powell, 1988; L. A. James, & L. R. James, 1989; MacCallum, & Browne, 1993; McCardle, 1989; Parkerson, Lomax, Schiller, & Walberg, 1984; Shavelson, & Bolus, 1982). Предпринимаются и попытки унифицировать и упростить процедуру моделирования структурными уравнениями (например, Jöreskog, & Sörbom, 1993).

Вклад когнитивной психологии. Семидесятые годы свидетельствовали о сближении между экспериментальной психологией и психометрией, которое начинает приносить плоды, крайне важные для понимания конструктов, оцениваемых с помощью тестов интеллекта и других широко определяемых способностей (Ronning, Glover, Conoley, & Witt, 1987; R. E. Snow, & Lohman, 1989). Еще в 1950-е гг. когнитивные психологи стали применять понятия теории информации при изучения процессов решения задач человеком (*human problem-solving*). Некоторые исследователи создавали компьютерные программы, которые осуществляли эти процессы и, таким образом, моделировали мышление человека. Можно написать программы, моделирующие деятельность людей на разных уровнях умения, и, располагая такими программами, предсказывать число и виды допускаемых ошибок, а также время, необходимое для различных реакций. При разработке программы исследователь обычно начинает с анализа задачи, в котором может использовать данные, полученные с помощью методик самонаблюдения, «размышления вслух» или каких-то более тонких методов. Сравнивая действия компьютера с действиями детей и взрослых (или с действиями экспертов и неопытных специалистов) при решении одной и той же задачи, исследователи получают возможность проверить свои гипотезы относительно того, что действительно люди делают при выполнении определенных заданий. Примеры задач, исследовавшихся этими методами, включают обычные головоломки, логические, шахматные, алгебраические (доказательство тождеств) и физические задачи, а также задачи медицинской диагностики (Chi, Glaser, & Farr, 1988; J. H. Larkin, McDermott, Simon, & Simon, 1980a, 1980 b; Newell, & Simon, 1972; Simon, 1976).

Выявленные в этих исследованиях переменные включают процессы (процедурные умения и навыки) и декларативные знания (факты и сведения). Когнитивные модели точно определяют интеллектуальные процессы, используемые при выполнении задания, способ организации этих процессов, запас релевантных знаний и то, как эти знания представлены в памяти и как они извлекаются из нее при необходимости. Все большее внимание уделяется и тому, что получило название исполнительного процесса или метапознания, относящегося к осуществляемому индивидуумом контролю за собственным выбором процессов, репрезентаций и стратегий для выполнения определенного задания. В 1970-х гг. некоторые когнитивные психологи начали применять эти методы анализа задач и компьютерного моделирования в поисковых исследованиях того, что же все-таки измеряют тесты интеллекта. Разные исследователи пыта-

лись подступиться к этой проблеме с разных сторон (см. Resnick, 1976; Sternberg, 1981, 1984, 1985b). Сумма полученных в таких исследованиях результатов постепенно переходит в важные достижения в области конструирования и применения тестов.

Следствия исследований в когнитивной психологии для валидизации конструктов особенно ясно показаны в работах Эмбретсона (Embretson, 1983, 1986, 1995a). Отмечая ограниченность традиционного подхода к валидизации конструктов, Эмбретсон предложил учитывать два принципиальных аспекта установления валидности теста: 1) репрезентацию конструкта и 2) номотетический диапазон. Традиционный подход к установлению конструктной валидности сосредоточивался полностью на втором аспекте, т. е. на определении номотетического диапазона теста. В этом случае рассматриваются связи результатов теста внутри «номотетической сети» других переменных. Такие связи обычно изучают путем вычисления корреляций тестовых показателей с другими мерами, включая результаты критериальной деятельности и иные жизненные показатели.

С другой стороны, цель репрезентации конструкта состоит в том, чтобы установить специфические компоненты процесса обработки информации и запасы знаний, которые нужны для выполнения задач, поставленных перед испытуемыми в заданиях теста. При проведении такого анализа можно применять метод *декомпозиции задачи (task decomposition)*.[1] Примеры возможных приемов включают манипулирование сложностью задачи, предъявление неполных задач или снабжение подсказками, изменяющими требования задачи. Для оценки вклада различных компонентов ответной реакции тестируемых в выполнение задания были разработаны специальные математические модели. Другим широко используемым методом для когнитивного анализа задачи является *анализ протоколов (protocol analysis)* (Ericsson, 1987; Ericsson, & Simon, 1993; van Someren, Barnard, & Sandberg, 1994). Этот метод предполагает инструкцию «думать вслух» при выполнении задания или во время решения задачи. Круг используемых заданий и задач довольно широк: от умножения в уме двух заданных чисел, припоминания деталей прошлого события или обнаружения причины неисправности оборудования до ответов на последовательность заданий теста способностей. Побочным продуктом этого метода является возможное обнаружение того, что одно и то же задание теста может вызывать совершенно разные когнитивные процессы у респондентов, различающихся по биографическим данным.

Какой вывод можно сделать на сегодняшний день в отношении вклада когнитивной психологии в развитие методов валидизации конструктов? Несмотря на то что сам информационный подход находится в стадии становления, он дал ряд эвристических концепций и руководящих принципов для организации дальнейших исследований в области валидизации тестов. Один из важнейших вкладов этого подхода — привлечение внимания к *процессуальной стороне ответов на задания тестов (response processes)*, в противоположность сосредоточению на конечных продуктах мышления в традиционных психометрических исследованиях. Анализ выполнения теста с точки зрения специфических когнитивных процессов определенно должен улучшить и расширить наше понимания того, что в действительности измеряют тесты. Кроме того, компонентный анализ индивидуального выполнения заданий теста на уровне элементарных процессов должен, в конечном счете, сделать возможным выявление

---

[1] Подробнее об этом см. Butterfield, Nielsen, Tangen, & Richardson (1985), Embretson (1985b) и Sternberg (1977, 1980).

слабых и сильных сторон каждого тестируемого и тем самым повысить значимость и привлекательность диагностического использования тестов (Embretson, 1987, 1994; Estes, 1974; Pellegrino, & Glaser, 1979; Sternberg, & Weil, 1980). А это, в свою очередь, должно облегчить приспособление программ обучения к потребностям каждого конкретного человека. Подводя итог, отношения между психометрическим и когнитивным подходами можно охарактеризовать, во-первых, с точки зрения прикладных исследований и практики, как *комплементарные*. В данном случае каждый подход специфичен в том, что касается целей, задач и методов исследования. Во-вторых, с точки зрения фундаментальных исследований и теории, их отношения можно охарактеризовать как *реципрокные*. Каждый подход способствует прояснению и обогащению другого, а вместе они улучшают наше понимание интеллектуального поведения.

Общий обзор и интеграция понятий

Сравнение методов валидизации. Мы рассмотрели несколько способов постановки вопроса «Насколько валиден данный тест?» Чтобы четче выделить отличительные признаки разных методов установления валидности, применим каждый из них по очереди к тесту, состоящему из 50 систематизированных арифметических задач. В табл. 5–2 представлены 4 возможных способа использования этого теста и соответствующие им методы валидизации. Из таблицы видно, что выбор метода валидизации зависит от последующего использования тестовых показателей. Валидность одного и того же теста в зависимости от цели его применения должна устанавливаться разными способами. Если тест достижений используется для предсказания дальнейших успехов на более высоком уровне обучения, как в случае отбора старшеклассников при их приеме в колледж, то валидность этого теста нужно оценивать относитель-

Таблица 5–2

Валидизация одного арифметического теста для разных целей

Цель тестирования	Иллюстративный вопрос	Доказательство валидности
Использование в качестве теста достижений по арифметике в начальной школе	Чему Дик научился на сегодняшний день?	Описание содержания
Использование в качестве теста способностей для предсказания успеваемости по математике в средней школе	Как хорошо будет учиться Джейн в дальнейшем?	Предсказание критерия (временное)
Использование в качестве способа диагностики трудностей в обучении	Указывает ли выполнение теста Биллом на какие-то специфические трудности в обучении?	Предсказание критерия (текущее)
Использование в качестве средства измерения количественных рассуждений	Как показатель Элен связан с другими показателями ее способности к логическим рассуждениям?	Идентификация конструкта

но такого критерия, как успешность обучения в колледже, а не относительно содержания данного школьного курса.

Инклюзивность валидизации конструктов. Примеры в табл. 5–2 подчеркивают различия между разными типами методов валидизации. Дальнейшее изучение этих методов, однако, показывает, что устанавливаемые с их помощью содержательная, прогностическая и конструктная валидности не соответствуют строго разграниченным или логически скоординированными категориям. Напротив, конструктная валидность — это широкое понятие, включающее другие типы валидности. Все обсуждавшиеся выше конкретные способы анализа содержания и оценки связей показателей теста с критерием можно было бы, кроме того, отнести и к категории способов идентификации конструкта. Например, корреляции теста механических способностей с успешностью обучения на специализированных курсах и с выполнением различного рода работ позволяет нам лучше понять конструкт, измеряемый данным тестом. Идентификацию этого конструкта можно дополнительно подкрепить сравнением показателей контрастных групп успешно и неуспешно работающих.

Валидность относительно разнообразных практических критериев обычно приводится в руководствах к тесту с тем, чтобы будущему пользователю легче было понять, что измеряет тест. Даже не будучи заинтересован в предсказании какого-либо из использованных конкретных критериев, он по их списку сможет составить себе представление об области поведения, выборочно проверяемой данным тестом. Если мы разовьем эту мысль немного дальше, то увидим, что всякое использование теста и любое истолкование тестовых показателей предполагает наличие конструктной валидности, — факт, который получает все большее признание (J. P. Campbell, 1990a; Guion, 1991; Messick, 1980b, 1988, 1989; Tenopyr, 1986). Поскольку тесты редко, если вообще когда-либо, используют в условиях, идентичных тем, в которых собирались данные для их валидизации, это неизбежно предполагает некоторую степень обобщаемости результатов. Смысл, вкладываемый в тестовые показатели при их интерпретации, всегда опирается на конструкты, которые могут сильно различаться по ширине обобщения на области поведения, популяции и условия.

Мессик (Messick, 1980b, 1989) приводит убедительные аргументы в пользу того, чтобы сохранить термин «валидность» (*validity*), коль скоро им обозначается обоснованность смысловой интерпретации теста, только за конструктной валидностью. Другим методам обоснования теста, с которыми традиционно связывался этот термин, считает Мессик, следует подобрать более точно описывающие их сущность названия. И тогда содержательную валидность можно было бы заменить на «содержательную релевантность» (*content relevance*) и «содержательное покрытие» (*content coverage*) — для спецификации и репрезентативности содержания теста относительно проверяемой предметной (или поведенческой) области соответственно. А критериальную валидность — заменить на «прогностическую полезность» (*predictive utility*) и «диагностическую полезность», чтобы эти термины соответствовали прогностической и текущей валидизации. Эти более точные, в плане описания, обозначения несомненно способствуют лучшему пониманию того, что в действительности достигается различными методами валидизации. Тем не менее выделение различных типов валидизации полезно в качестве дополнительных опознавательных признаков тестов различного назначения. Поэтому об использованных типах валидизации следует сообщать в руководствах к тестам в легкоузнаваемой форме.

С другой стороны, даже когда непосредственная прикладная задача направлена на описание содержания (как в образовательном тестировании) или на предсказание критерия (как в профотборе), использование конструктов подходящей широты эффективнее применения мер конкретного выполнения теста. Исследования используемых в тестировании критериев делают все более очевидным тот факт, что и меры критерия и показатели теста можно более эффективно выразить в виде пары согласованных конструктов. Более того, изучение причинных отношений между конструктами, как при моделировании структурными уравнениями, получает признание в качестве важного вклада в понимание того, как и почему работают тесты.[1]

Валидизация в процессе конструирования теста. Все шире признается, что разработка валидного теста требует применения многих методов, используемых последовательно, на разных этапах конструирования теста (Anastasi, 1986a; Guion, 1991; Jackson, 1970, 1973; N. G. Peterson et al., 1990). Таким образом валидность теста создается постепенно, начиная с первого шага в его разработке, а вовсе не на последних этапах, как при традиционной валидизации относительно критерия. Процесс валидизации начинается с формулирования детальных определений черты, свойства или конструкта на основе психологической теории, предшествующих исследований или систематического наблюдения и анализа релевантной области поведения. Затем, в соответствии с определениями конструкта, готовят задания теста. За этим следует их эмпирический анализ, с отбором наиболее эффективных, или валидных, заданий из исходной совокупности. Далее могут проводиться различные виды внутреннего анализа, включая статистический анализ кластеров заданий или субтестов. Заключительный этап включает в себя валидизацию различных показателей и их интерпретируемых комбинаций посредством статистического анализа, но уже относительно внешних, реальных критериев.

Практически любые сведения, собранные в процессе разработки или использования теста, имеют отношение к его валидности и могут оказаться полезными. Данные о внутренней согласованности и ретестовой надежности, несомненно, помогают определить однородность конструкта и его временну́ю устойчивость. Нормы могут способствовать дополнительной детализации описания конструкта, особенно если они включают нормативные данные для подгрупп, сформированных по возрасту, полу или другим демографическим переменным, влияющим на биографию конкретного человека и тем самым на результаты теста. Кроме того, после всех испытаний теста и получения разрешения на его практическое использование смысловая интерпретация его показателей может уточняться и обогащаться благодаря постепенному накоплению клинических наблюдений и выполнению специальных исследовательских проектов.[2]

Индивидуальные и социальные последствия тестирования. Некоторые психометристы предлагали включить в понятие валидности теста дополнительный признак, а именно *последствия (consequences)* тестирования для конкретных людей и для общества в целом. Известным сторонником такого расширения понятия валидности является Мессик (Messick, 1980b, 1988, 1989, 1995). Особо выделяются непродуманные

[1] Пример возможного применения этих более тонких методов валидизации тестов можно найти в L. A. King & D. W. King (1990).

[2] Об удачном применении этой комплексной модели валидизации теста см. Elliott (1990b, chap. 9).

заранее последствия целевого применения тестов, которое может причинить вред отдельным лицам и членам определенных этнических или других групп с отличающийся от большинства историей жизни. Превосходный анализ проблем согласования различных целей и ценностей при оценивании претендентов на рабочие места иллюстрируется отчетом временно созванного Национальным научно-исследовательским советом (*National Research Council*) комитета экспертов, который с необычайной основательностью изучил эту ситуацию (Hartigan, & Wigdor, 1989 — см. особенно chaps. 13 и 14).

Этические и социальные последствия использования тестов бесспорно требуют самого широкого внимания. Некоторое ознакомление с этими проблемами дает глава 18. Их более специальные аспекты рассматриваются в главе 6, в связи с вопросом «необъективности тестов» (*test bias*). Однако, как отмечают другие психометристы (например, Cole, & Moss, 1989), включение этих вопросов в понятие валидности вряд ли будет самым эффективным способом их разрешения. На них невозможно ответить, опираясь только на эмпирические данные и статистический анализ. Да и вряд ли следует маскировать привлекательные для нас ценности статистическими манипуляциями. Эти вопросы нужно открыто формулировать и обсуждать как самостоятельную, объективную цель, рассматриваемую в дополнение к сугубо эмпирической и статистически доказанной валидности использования конкретного теста. Взвешенное решение, касающееся согласования конфликтующих целей, достигается методами, пригодными для преобразования систем ценностей (Mullen, & Roth, 1991; Zeichmeister, & Johnson, 1992).[1] Такие методы требуют специального разбирательства, систематических дискуссий, разрешения конфликтов и достижения компромиссов, причем должно быть обеспечено соразмерное представительство сторонников различных систем ценностей. Объединение эмпирических, статистически подкрепляемых процедур определения валидности с оцениванием социальных и этических последствий применения конкретного теста только затрудняет и затемняет решение.

Один вывод, который напрашивается при рассмотрении этой трудной и важной проблемы, — дополнительное подтверждение главной роли пользователя тестов, о чем уже говорилось в главе 1. Когда требуется переоценка ценностей, особенно в индивидуальных случаях, на пользователей тестов возлагается еще большая ответственность, ибо они могут контролировать последствия тестирования и при выборе подходящих тестов, и при интерпретации результатов. Толерантность к широкому спектру ценностей и социальная чувствительность пользователя могут в значительной мере способствовать правильному использованию тестов, причем не только с научной, но и с этической точки зрения.[2]

---

[1] См. также Arkes (1993), где эта проблема освещается более широко.

[2] Попутно можно отметить, что новый подход к психологии в целом предполагает построение «дискурсивной психологии», в которой проблемы изучаются как посредством их обсуждения между людьми в повседневной жизни, так и традиционными экспериментальными методами (см., например, Harré & Stearns, 1995; J. Smith, Harré, & Van Langenhove, 1995).

6 ВАЛИДНОСТЬ: ИЗМЕРЕНИЕ И ИНТЕРПРЕТАЦИЯ

Глава 5 была посвящена рассмотрению понятий валидности и источников данных валидизации тестов. В этой главе обсуждаются способы выражения валидности в количественной форме и интерпретация ее соответствующих числовых оценок. Пользователи напрямую сталкиваются с валидностью на одном из двух или на обоих этапах работы с тестом. Первый раз, оценивая пригодность теста для своих целей, они изучают данные о валидности, приведенные в руководстве к тесту или в других доступных источниках. На основе такой информации они получают предварительное представление о том, какие психологические функции тест измеряет, и оценивают, имеют ли эти функции отношение к предполагаемому использованию теста. В сущности, когда пользователи опираются в своей оценке только на опубликованные данные о валидности теста, они имеют дело с конструктной валидностью, независимо от конкретных методов сбора таких данных. Как уже отмечалось в главе 5, приводимые в опубликованных исследованиях критерии нельзя считать полностью идентичными тем, которые пользователи теста собираются прогнозировать. Даже одноименные должности на двух различных предприятиях редко совпадают по своим обязанностям, точно так же, как два курса английского языка, преподаваемые в разных колледжах первокурсникам, могут значительно отличаться друг от друга. Следовательно, какая-то степень обобщения валидности предполагается самим фактом выбора теста.

Ввиду различий в потребностях тестирования и в выводах, которые предполагается делать из тестовых показателей, у некоторых пользователей может появиться желание проверить валидность выбираемого теста относительно локальных критериев. Даже если опубликованные данные явно указывают на высокую валидность теста в определенной ситуации, ее прямое подтверждение, когда это технически возможно, никогда не будет лишним. Определение валидности относительно конкретных локальных критериев представляет собой второй этап в работе пользователей, когда им приходится иметь дело с валидностью теста. Методы, рассматриваемые в этой главе, имеют непосредственное отношение к анализу данных валидизации, получаемых самим пользователем теста, но они (по крайней мере, большая их часть) также полезны для понимания и интерпретации сведений о валидности, приводимых в руководствах к тестам.

Коэффициент валидности и ошибка оценки

Измерение соотношения. Коэффициент валидности выражает величину корреляции между показателем теста и мерой критерия. Этот коэффициент позволяет характеризовать валидность единственным числовым показателем, и поэтому его часто приводят в руководствах к тестам, указывая его величину для каждого из использованных критериев. Данные, по которым вычисляется коэффициент валидности, могут к тому же быть представлены в виде таблицы ожидаемых результатов или диаграммы ожидаемого отсева (см. главу 3). Собственно говоря, такие таблицы и диаграммы — наглядные иллюстрации того, что коэффициент валидности означает для тестируемого. Напомним, что в таблицах ожидаемых результатов приводятся вероятности достижения определенного уровня выполнения критериальной деятельности испытуемым, получившим определенный показатель по данному тесту. Например, с помощью табл. 3–6, зная показатель ученика по тесту числового рассуждения из батареи Дифференциальных тестов способностей (*DAT*), можно определить вероятность получения им той или иной оценки по математике в 7-м классе. Для тех же данных коэффициент валидности составляет 0,60. Если, как в приведенном примере, тестовая и критериальная переменные являются непрерывными, то применим уже знакомый нам коэффициент корреляции произведения моментов Пирсона. Если же исходные данные выражены в иной форме (скажем, при использовании дихотомического критерия «выполнено—невыполнено» — см. рис. 3–7), вычисляются другие виды коэффициентов корреляции. Соответствующие вычислительные процедуры можно найти в любом типовом учебнике по статистике.

Условия, влияющие на величину коэффициентов валидности. Как и в случае с надежностью, важно точно определять *характер группы,* на которой вычисляется коэффициент валидности теста. Один и тот же тест может измерять различные функции, если его дать лицам разного возраста, пола, уровня образования, рода занятий и т. д. Люди с разным жизненным, учебным и профессиональным опытом могут, например, воспользоваться разными методами для решения одной и той же тестовой задачи. Следовательно, тест может обладать высокой валидностью относительно заданного критерия в одной популяции и низкой или нулевой валидностью — в другой. Или, скажем, оказаться валидной мерой разных функций в двух популяциях. Поэтому в технических руководствах к тестам, предназначенным для работы с разнотипными популяциями, следует приводить соответствующие данные о популяционной обобщаемости (*population generalizability*). Кроме того, когда имеет место значительная внутрипопуляционная вариация тестовых показателей, коэффициент валидности теста может заметно различаться в разных частях диапазона показателей и должен проверяться в соответствующих подгруппах (R. Lee, & Foley, 1986).

Вопрос *неоднородности выборки* имеет для измерения валидности такое же значение, как и для измерения надежности, поскольку обе характеристики обычно приводятся в виде коэффициентов корреляции. Напомним, что при прочих равных условиях чем шире размах распределения показателей, тем выше будет корреляция. Это обстоятельство необходимо иметь в виду при интерпретации коэффициентов валидности, приводимых в руководствах к тестам.

Специфическая проблема, присущая многим выборкам валидизации, связана с *предотбором (preselection)*. Например, новый тест, валидизируемый для целей профотбора,

может проводиться на группе недавно нанятых работников, в отношении которых со временем будут доступны такие меры критерия, как эффективность труда. Вполне вероятно, однако, что эти работники представляют собой верхнюю (лучшую) часть выборки из всех тех, кто хотел поступить на эту работу. Поэтому нижний конец распределения тестовых показателей и критериальных мер в такой выборке окажется обрезанным. Эффектом такого предотбора, естественно, будет снижение коэффициента валидности. При последующем использовании теста, когда его будут проводить со всеми поступающими на работу в целях их отбора, можно ожидать некоторого повышения его валидности.

Коэффициенты валидности могут также измениться через какое-то время вследствие изменения норм отбора. В качестве примера сравним коэффициенты валидности, полученные с интервалом в 30 лет при обследовании студентов Йельского университета (Burnham, 1965). Определялась корреляция между прогнозирующим показателем, основанным на тестах Совета колледжей, и успеваемостью в старших классах, с одной стороны, и средним баллом первокурсника — с другой. Оказалось, что за 30 лет корреляция снизилась с 0,71 до 0,52. Анализ соответствующих двумерных распределений данных легко выявил причину этого снижения. Дело в том, что в связи с повысившимися требованиями при приеме в колледж группа студентов во втором случае стала более однородной, чем в первом, по отношению как к прогнозирующему показателю, так и к мерам критерия. Отсюда и падение корреляции, несмотря на то что точность прогноза успеваемости в колледже осталась, в общем, прежней. Иными словами, наблюдаемое снижение корреляции вовсе *не* свидетельствовало о том, что прогнозирующие показатели стали менее валидными, чем 30 лет назад. А ведь именно к такому выводу можно было бы прийти, упустив из виду различия в однородности групп.

Для правильной интерпретации коэффициента валидности следует принимать во внимание и *форму связи* между тестом и критерием. Вычисление пирсоновского коэффициента корреляции предполагает, что эта связь линейна и остается неизменной во всем диапазоне распределения. Исследование связи тестовых показателей с выполнением работы показало, что эти условия, в общем, выполняются (Coward, & Sackett, 1990; Hawk, 1970). Все же особые обстоятельства могут изменять характер этой связи, и пользователю теста следует быть всегда готовым к такому повороту событий. Пусть для выполнения некоторой работы требуется лишь минимальный уровень понимания читаемого, достаточный для прочтения инструкций, названий и т. д. Но как только этот минимальный уровень превзойден, то от дальнейшего развития данного умения успешность выполнения работы уже не зависит, т. е. между тестом и выполнением работы существуют нелинейные отношения. Изучение двумерного распределения или диаграммы рассеяния, построенной по показателям теста на понимание читаемого и мерам критерия, в этом случае показало бы, что уровень выполнения работы растет, пока умение понимать читаемое не достигает требуемой степени, после чего он остается примерно тем же. Следовательно, точки на диаграмме группируются вокруг кривой, а не прямой линии.

В других случаях линия наилучшего соответствия может быть и прямой, но точки, соответствующие индивидуальным данным, могут отклоняться от нее в верхнем конце шкалы больше, чем в нижнем. Предположим, что успешное выполнение теста академических способностей — необходимое, но не достаточное условие для успешного завершения некоторого учебного курса. Это значит, что учащиеся с низкими показа-

телями по данному тесту получат скорее всего неудовлетворительные оценки, тогда как среди учащихся с высокими показателями одни получат положительные оценки, а другие, из-за недостаточной мотивации, отсутствия интереса или других неблагоприятных условий, не сдадут экзамена. В этой ситуации будет наблюдаться большая вариативность выполнения критериальной деятельности у учащихся с высокими тестовыми показателями, чем с низкими. Такое условие в двумерном распределении называется гетероскедастичностью.[1] Пирсоновская корреляция предполагает гомоскедастичность, т. е. одинаковую вариабельность во всем диапазоне двумерного распределения. В приведенном примере двумерное распределение было бы веерообразным — широким в верхнем конце и узким в нижнем. Уже визуального анализа двумерного распределения обычно бывает достаточно для установления характера связи между тестом и критерием. Таблицы ожидаемых результатов и диаграммы ожидаемого отсева также правильно показывают относительную эффективность теста на разных уровнях.

Величина коэффициента валидности. Какова должна быть величина коэффициента валидности? На этот вопрос нет единого ответа, так как при интерпретации коэффициента валидности нужно учитывать ряд сопутствующих обстоятельств. Разумеется, корреляция должна быть достаточно высокой для того, чтобы быть *статистически значимой* на приемлемом уровне, таком как 0,01 или 0,05 (см. главу 4). Иными словами, прежде чем делать какие-либо выводы о валидности теста, нужно иметь обоснованную уверенность в том, что полученный коэффициент валидности не появился в результате случайных колебаний выборки из генеральной совокупности с нулевой корреляцией.

Установив значимую корреляцию между тестовыми показателями и критерием, необходимо еще оценить ее величину в аспекте тех целей, ради которых и создавался данный тест. Если мы собираемся предсказывать точное значение критериального показателя у конкретных лиц (скажем, средний балл студента в колледже), коэффициент валидности можно интерпретировать исходя из *стандартной ошибки оценки* (*standard error of estimate*, или сокращенно, SE_{est}), которая аналогична ошибке измерения, обсуждавшейся в связи с надежностью. Напомним, что ошибка измерения указывает допустимый предел возможной ошибки индивидуального показателя в результате ненадежности теста. Аналогично этому, ошибка оценки указывает допустимый предел возможной ошибки прогнозируемой величины индивидуального критериального показателя в результате недостаточной валидности теста.

Ошибка оценки вычисляется по следующей формуле:

$$SE_{est} = SD_y\sqrt{1-r_{xy}^2},$$

где r_{xy}^2 — квадрат коэффициента валидности и SD_y — стандартное отклонение критериального показателя. Заметим, что при полной валидности ($r_{xy} = 1{,}00$) ошибка оценки была бы равна нулю. С другой стороны, если валидность теста равна нулю, то ошибка оценки достигает величины стандартного отклонения распределения критерия $\left(SE_{est} = SD_y\sqrt{1-0} = SD_y\right)$. При этих условиях вероятность правильного прогноза не

[1] Термины «гомоскедастичность» и «гетероскедастичность» (букв. «одинаковая рассеянность» и «неодинаковая рассеянность» соответственно) введены в статистику А. А. Чупровым. — *Примеч. науч. ред.*

превышает вероятности случайного угадывания, и диапазон ошибки предсказания равен ширине распределения критериальных показателей. Между этими двумя пределами и будут заключаться ошибки оценки, соответствующие тестам с варьирующей валидностью.

Обращаясь к формуле для SE_{est}, покажем, что выражение $\sqrt{1-r_{xy}^2}$ позволяет определить величину ошибки оценки *относительно ошибки простого угадывания* (т. е. при нулевой валидности). Иными словами, если $\sqrt{1-r_{xy}^2} = 1{,}00$, то ошибка оценки столь же велика, как и при случайном угадывании критериального показателя у конкретного испытуемого. Использование такого теста не дало бы нам никакого выигрыша в точности предсказания. Если же коэффициент валидности равен 0,80, то $\sqrt{1-r_{xy}^2} = 0{,}60$, и максимальная ошибка составляет 60 % от величины той, которая была бы при случайном угадывании. Выражаясь иначе, использование этого теста позволяет нам предсказывать индивидуальные результаты в критериальной деятельности с пределом ошибки, который на 40 % меньше, чем в случае угадывания.

Может показаться, что даже при такой необычайно высокой валидности, как 0,80, ошибка предсказываемых показателей довольно значительна. Если бы главной функцией психологических тестов было предсказание точного положения индивидуума в критериальном распределении, такая перспектива выглядела бы совершенно обескураживающей. Когда мы рассматриваем тесты в аспекте ошибки оценки, большинство из них представляются не особенно эффективными. Однако чаще всего при тестировании нет необходимости предсказывать точный результат критериальной деятельности каждого обследуемого человека, но требуется лишь определить, кто из них превзойдет некоторый минимальный стандарт выполнения, или критический показатель выбранной в качестве критерия деятельности. Каковы шансы у Мери Грин закончить медицинское училище, у Тома Хиггинса усвоить курс вычислительной математики, а у Беверли Брюса преуспеть в качестве астронавта? Кто из поступающих на работу, скорее всего, будет хорошим клерком, страховым агентом, механиком? Такая информация полезна не только для профотбора, но и для профориентации. Например, студенту полезно и выгодно знать, что у него хорошие шансы благополучно окончить юридический факультет, даже если мы не можем с уверенностью сказать, будет ли его средний балл 74 или 81.

Тест может заметно повысить свою предсказуемостную эффективность, если для него будет установлена *любая* значимая корреляция с критерием, какой бы низкой она ни была. При некоторых обстоятельствах валидность порядка 0,20–0,30 уже оправдывает включение теста в программу отбора. Для многих целей тестирования оценивание тестов с точки зрения их стандартной ошибки оценки является неоправданно строгим. В большинстве случаев должны применяться другие способы оценивания тестов, те, которые бы учитывали типы решений, принимаемых на основе их результатов. О некоторых из них пойдет речь в следующем разделе.

Валидность теста и теория принятия решений

Основной подход. Предположим, 100 человек, поступающих на работу, выполнили тест способностей и по прошествии какого-то времени были оценены их успехи в выполнении своих обязанностей. На рис. 6–1 изображено соответствующее двумерное распределение тестовых показателей и мер успешного выполнения работы. Корре-



Рис. 6–1. Прирост доли «успешных работников» вследствие использования теста отбора

ляция между этими двумя переменными чуть ниже 0,70. Необходимый минимум выполнения работы, или критический показатель, отмечен на диаграмме жирной горизонтальной линией. Сорок случаев, лежащих ниже этой линии, соответствуют числу людей, не справившихся с работой, а 60 случаев выше нее — числу успешно работающих. Если на работу принимаются все 100 человек, подавших заявление, то, следовательно, 60 % справятся с ней. Подобным же образом, если бы меньшее число работников нанималось наугад, безотносительно к результатам тестирования, доля успешно справившихся с работой была бы, вероятно, близка к 60 %. Предположим, однако, что тестовые показатели используют для отбора из 100 претендентов 45 наиболее перспективных работников (коэффициент отбора = 0,45). В таком случае были бы выбраны 45 человек, попадающие в область справа от жирной вертикальной линии. На диаграмме видно, что из этих 45 человек 7 попадают ниже жирной горизонтальной линии, т. е. в разряд несправившихся с работой, и составляют долю *ошибочно принятых*, а 38 человек — в разряд успешных работников. Процент успешно справившихся с работой теперь уже равен не 60, а 84 (т. е. 38 / 45 = 0,84). Это увеличение обусловлено применением теста в качестве инструмента отбора. Заметим, что ошибками показателя предсказываемого критерия, не влияющими на принятие решение, можно пренебречь. Селективную эффективность теста будут снижать только те ошибки предсказания, которые ведут к пересечению линии критического показателя и, следовательно, к помещению индивидуума в ошибочную категорию.

Для полной оценки эффективности теста как инструмента отбора необходимо также изучить другую категорию случаев, отображенную на рис. 6–1. Это категория *ошибочно непринятых*, включающая 22 человека, у которых показатели по тесту ниже критического уровня, а показатели критериальной деятельности выше такового.

Исходя из полученных данных, можно приблизительно оценить, что 22 % всей выборки претендентов на получение работы, являясь потенциально успешными работниками, будут потеряны в том случае, когда данный тест применяется в качестве инструмента отбора с выбранным таким образом критическим показателем. Устанавливая уровень критического показателя по тесту, следует учитывать процент случаев ошибочного отказа в приеме, а также процент успешных и неуспешных работников в группе отобранных. В определенных ситуациях уровень устанавливаемого критического показателя должен быть достаточно высоким, чтобы почти полностью исключить возможные неудачи. Это необходимо, когда характер работы таков, что недостаточно квалифицированный работник может нанести серьезный ущерб или вред. В качестве примера здесь уместно указать на отбор пилотов гражданской авиации. При других обстоятельствах бывает важнее нанять как можно больше квалифицированных работников, идя на риск принять и больше неспособных к данному роду деятельности. В последнем случае число ошибочных отказов сокращается за счет выбора более низкого уровня критического показателя. К другим факторам, которые обычно влияют на уровень критического показателя, относятся число претендентов, количество вакансий и сроки, в которые эти вакансии необходимо заполнить.[1]

Во многих кадровых решениях коэффициент отбора определяется практическими требованиями конкретной ситуации. В одних случаях соотношение спроса и предложения обусловливает, например, прием 40 %, а в других — 75 % претендентов (с лучшими показателями, разумеется). Если коэффициент отбора не диктуется внешними обстоятельствами, то критический показатель по тесту может устанавливаться на уровне, обеспечивающем наилучшую дифференциацию двух групп по критериальной деятельности. Приблизительно это можно сделать, сравнивая распределение показателей теста в группах «успешных» и «неуспешных» работников. Разработаны и более точные математические методы определения оптимального уровня критических показателей по тесту (Darlington, & Stauffer, 1966; I. Guttman, & Raju, 1965; Jaeger, 1989; Livingston, & Zieky, 1982; Martin, & Raju, 1992; Rorer, Hoffman, & Hsieh, 1966). Эти методы позволяют учитывать другие релевантные параметры, такие как относительная серьезность ошибочных отказов и необоснованного приема на работу. Однако поскольку такие оценки включаются в реализацию этих методов, постольку на определенном этапе все равно возникает потребность в человеческих, а значит и субъективных, суждениях.

На языке теории принятия решений, представленный на рис. 6–1 пример иллюстрирует простую *стратегию* отбора претендентов. В более широком смысле, стратегия — это способ использования информации для выработки решения в отношении определенного круга лиц. В данном случае стратегия состоит в приеме 45 человек с самыми высокими тестовыми показателями. Увеличение доли успешно справляющихся со своей работой лиц с 60 до 84 % могло бы послужить основанием для оценивания чистой выгоды от использования теста.

Теория статистических решений была разработана А. Вальдом (Wald, 1950) применительно к решениям, принимаемым, в основном, при выборочном контроле качества массовой продукции. Многие из ее выводов и следствий для конструирования и интерпретации психологических тестов систематически развивали Кронбах и Глесер

---

[1] Сходные вопросы уже рассматривались под другим углом зрения при предварительном обсуждении критических показателей в главе 3.

(Cronbach, & Gleser, 1965). В сущности, теория решений представляет собой попытку придать процессу принятия решения математическую форму, с тем чтобы использовать имеющуюся информацию для выработки в конкретных обстоятельствах наиболее эффективных решений. Основные понятия теории принятия решений оказываются полезными для переформулирования и прояснения ряда связанных с тестами вопросов. Некоторые из них были введены в тестирование еще до того, как был разработан формальный аппарат теории статистических решений, и позднее были признаны соответствующими ее аппарату.

Предсказание результатов. Своего рода предшественником теории принятия решений в психологическом тестировании явились таблицы Тейлора—Расселла (Н. С. Taylor, & Russell, 1939), позволявшие определить чистый выигрыш в точности отбора за счет использования теста. Для работы с таблицами нужно знать коэффициент валид-

Таблица 6–1

Доля «успешных работников», на которую можно рассчитывать при заданном коэффициенте отбора и заданной валидности используемого теста (базисная норма = 0,60)

Валидность	Коэффициент отбора										
	0,05	0,10	0,20	0,30	0,40	0,50	0,60	0,70	0,80	0,90	0,95
0,00	0,60	0,60	0,60	0,60	0,60	0,60	0,60	0,60	0,60	0,60	0,60
0,05	0,64	0,63	0,63	0,62	0,62	0,62	0,61	0,61	0,61	0,60	0,60
0,10	0,68	0,67	0,65	0,64	0,64	0,63	0,63	0,62	0,61	0,61	0,60
0,15	0,71	0,70	0,68	0,67	0,66	0,65	0,64	0,63	0,62	0,61	0,61
0,20	0,75	0,73	0,71	0,69	0,67	0,66	0,65	0,64	0,63	0,62	0,61
0,25	0,78	0,76	0,73	0,71	0,69	0,68	0,66	0,65	0,63	0,62	0,61
0,30	0,82	0,79	0,76	0,73	0,71	0,69	0,68	0,66	0,64	0,62	0,61
0,35	0,85	0,82	0,78	0,75	0,73	0,71	0,69	0,67	0,65	0,63	0,62
0,40	0,88	0,85	0,81	0,78	0,75	0,73	0,70	0,68	0,66	0,63	0,62
0,45	0,90	0,87	0,83	0,80	0,77	0,74	0,72	0,69	0,66	0,64	0,62
0,50	0,93	0,90	0,86	0,82	0,79	0,76	0,73	0,70	0,67	0,64	0,62
0,55	0,95	0,92	0,88	0,84	0,81	0,78	0,75	0,71	0,68	0,64	0,62
0,60	0,96	0,94	0,90	0,87	0,83	0,80	0,76	0,73	0,69	0,65	0,63
0,65	0,98	0,96	0,92	0,89	0,85	0,82	0,78	0,74	0,70	0,65	0,63
0,70	0,99	0,97	0,94	0,91	0,87	0,84	0,80	0,75	0,71	0,66	0,63
0,75	0,99	0,99	0,96	0,93	0,90	0,86	0,81	0,77	0,71	0,66	0,63
0,80	1,00	0,99	0,98	0,95	0,92	0,88	0,83	0,78	0,72	0,66	0,63
0,85	1,00	1,00	0,99	0,97	0,95	0,91	0,86	0,80	0,73	0,66	0,63
0,90	1,00	1,00	1,00	0,99	0,97	0,94	0,88	0,82	0,74	0,67	0,63
0,95	1,00	1,00	1,00	1,00	0,99	0,97	0,92	0,84	75	0,67	0,63
1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	0,86	0,75	0,67	0,63

Примечание. Полный набор таблиц можно найти в Н. С. Taylor & Russell (1939) и в McCormick & Ilgen (1980, Appendix В).

(Из Н. С. Taylor & Russell, 1939, p. 576)

ности теста, долю претендентов, которых необходимо принять (коэффициент отбора), и долю успешно справляющихся с обязанностями работников, отобранных без использования теста (базисную норму). Изменение любого из этих условий может повлиять на предсказуемостную эффективность теста.

В целях иллюстрации воспроизведена одна из таблиц Тейлора—Расселла (табл. 6–1). Данная таблица предназначена для использования с базисной нормой (процентом успешных работников, отобранных до использования теста), равной 0,60. Аналогичные таблицы составлены Тейлором и Расселом для других базисных норм. В верхней строке таблицы приведены различные значения коэффициента отбора, в крайнем левом столбце — коэффициенты валидности теста, а числа на пересечении каждой строки и столбца показывают долю успешных работников, отобранных с помощью тестирования. Разность между любым таким числом и базисной нормой (0,60) показывает прирост правильно отобранных работников за счет использования теста.

Очевидно, если коэффициент отбора равен 100 %, т. е. когда пришлось бы принимать на работу всех претендентов, ни один тест, какой бы высокой ни была его валидность, не улучшил бы качества отбора. Из табл. 6–1 видно, что при коэффициенте отбора, равном 0,95, даже абсолютно валидный тест ($r = 1{,}00$) повысил бы долю успешных работников только на 3 % (с 0,60 до 0,63). Напротив, если из поступающих нужно отобрать только 5 %, то тест с коэффициентом валидности, равным всего 0,30, может повысить процент удачно отбираемых работников с 60 до 82. Этот прирост с 60 до 82 % отражает *инкрементную валидность (incremental validity)* теста (Sechrest, 1963), или увеличение прогностической валидности, свойственной данному тесту. Инкрементная валидность показывает вклад теста в отбор лиц, которые в дальнейшем будут удовлетворять минимальным требованиям критериальной деятельности. При применении таблиц Тейлора—Расселла, валидность теста, разумеется, должна определяться на группе того же типа, которая использовалась для оценки базисной нормы. Иными словами, вклад теста не оценивается относительно случайного успеха, если только претендентов до этого не отбирали наугад, что весьма маловероятно. Если же претендентов отбирали на основе сведений о предыдущей работе, рекомендательных писем и результатов собеседования, то и вклад теста следует оценивать по тому, что он добавляет к таким методам отбора.

Инкрементная валидность, вытекающая из использования теста, зависит не только от коэффициента отбора, но и от базисной нормы. В рассматриваемой нами иллюстративной ситуации базисная норма указывает на долю успешных работников до момента внедрения теста в целях профотбора. В табл. 6–1 приведены ожидаемые результаты при базисной норме 0,60. В случае других базисных норм нам придется обратиться к другим, соответствующим таблицам в указанном источнике (H. C. Taylor, & Russell, 1939). Давайте рассмотрим пример, когда валидность теста равна 0,60, а коэффициент отбора — 40 %. Каков был бы вклад инкрементной валидности теста при этих условиях, если бы мы начали с базисной нормы в 50 %? И что изменилось бы, если бы мы перешли к таким крайним значениям базисной нормы, как 10 % и 90 %? Обращение к соответствующим таблицам Тейлора—Расселла показывает, что процент успешных работников повысился бы с 50 до 75 в первом случае, с 10 до 21 во втором и с 90 до 99 в третьем. Таким образом, увеличение доли успешных работников, которое можно приписать применению теста, составляет 25 % при базисной норме в 50 %, но только 11 % и 9 % при крайних значениях базисной нормы.

Поведение инкрементной валидности при базисных нормах, близких к нулю или единице, представляет особый интерес для клинической психологии, где базисная норма говорит о частоте патологических состояний, диагностируемых в обследуемой популяции (Buchwald, 1965; Cureton, 1957a; Meehl, & Rosen, 1955; J. S. Wiggins, 1973/1988). Например, если у 5 % помещаемых в клинику лиц имеется органическое поражение мозга, то базисная норма для данного диагноза в данной популяции будет равна 5 %. Хотя внедрение любого валидного теста повысит точность диагностики или прогноза, улучшение точности будет максимальным лишь тогда, когда базисные нормы близки к 50 %.

При низких базисных нормах, соответствующих редким патологическим состояниям, это улучшение может оказаться незначительным. В таких случаях использование теста нельзя будет считать оправданным, учитывая издержки, связанные с его проведением и обработкой результатов. В условиях клиники такие издержки включали бы время квалифицированного персонала, которое иначе можно было бы потратить на лечение дополнительных больных (Buchwald, 1965). Какое-то количество ложных положительных диагнозов (*false positives*), т. е. нормальных лиц, ошибочно отнесенных к той или иной патологии, еще более увеличило бы эти общие издержки в клинической ситуации.

Когда редкая патология настолько серьезна, что необходим срочный диагноз, тесты с умеренной валидностью можно использовать на раннем этапе последовательных диагностических решений. Например, всех пациентов можно обследовать с помощью легко проводимого теста с невысокой валидностью. Если устанавливается достаточно высокий критический показатель (высокие показатели в данном случае предпочтительней), то число ложных отрицательных диагнозов (*false negatives*) будет мало, а число ложных положительных диагнозов, напротив, велико. Последние затем могут быть выявлены при более интенсивном индивидуальном обследовании всех получивших положительный диагноз по тесту. Такой подход целесообразен, когда, например, имеющееся оборудование не позволяет проводить интенсивного индивидуального обследования всех пациентов.

Отношение валидности к продуктивности. Во многих практических ситуациях требуется оценить эффективность теста для профотбора не по проценту лиц, преодолевших «планку» минимальных требований к деятельности, а по предельной продуктивности труда отобранных с его помощью работников. Как реальный уровень квалификации работников (или выполнения ими критериальной деятельности), нанятых по результатам тестирования, сравнить с уровнем общей выборки кандидатов, которые могли бы быть приняты на работу без проведения данного теста? После появления работы Тейлора и Расселла некоторые исследователи заинтересовались этим вопросом. Брогден (Brogden, 1946b) первым показал, что ожидаемый прирост продуктивности прямо пропорционален валидности теста. Так, улучшение от применения теста с валидностью 0,50 составляет 50 % улучшения, ожидаемого при использовании абсолютно валидного теста.

Связь между валидностью теста и ожидаемым повышением критериальных достижений видна из табл. 6–2. Выражая критериальные показатели в виде стандартных показателей со средним, равным нулю, и $SD = 1$, эта таблица содержит ожидаемые средние критериальных показателей работников, отобранных при заданном коэффи-

циенте отбора с помощью теста, имеющего определенную валидность.[1] В этом контексте средняя базисная продуктивность, соответствующая деятельности работников, набранных без использования теста, приводится в колонке нулевой валидности. Использовать тест с нулевой валидностью — это все равно, что не использовать никаких тестов. Покажем, как пользоваться этой таблицей. Предположим, приему подлежат 20 % претендентов с самыми высокими показателями (коэффициент отбора 0,20), причем отбор производится с помощью теста, валидность которого равна 0,50. По табл. 6–2 находим, что средний критериальный показатель в отобранной группе превышает средний показатель базисной продуктивности на 0,7 *SD*. При том же коэффициенте отбора (0,20) и применении идеального теста (с коэффициентом валидности 1,00) средний критериальный показатель принятых на работу претендентов составил бы уже 1,40, т. е. оказался бы ровно в два раза выше, чем при использовании теста с валидностью 0,50. Подобная прямая линейная зависимость имеет место в пределах любой строки табл. 6–2. Например, при коэффициенте отбора 0,60 тест с валидностью 0,25 дает средний критериальный показатель 0,16, в то время как тест с валидностью 0,50 обеспечивает средний критериальный показатель 0,32. Опять-таки удвоение валидности ведет к удвоению показателя продуктивности.

Анализ продуктивности в связи с валидностью тестов, используемых для отбора кадров, был продолжен Шмидтом и его коллегами (Schmidt, Hunter, McKenzie, & Muldrow, 1979). Выбрав в качестве иллюстративного образца работу программиста в федеральном правительстве, эти исследователи оценили в долларовом эквиваленте повышение продуктивности в результате использования в течение года теста компьютерных способностей (*computer aptitude test*) (коэффициент валидности равен 0,76) при отборе наемных работников. Они получили свои оценки, применяя методы теории принятия решений к данным, имеющимся в распоряжении Службы управления кадрами США (*U. S. Office of Personnel Management*). Ожидаемая прибыль рассчитывалась для девяти коэффициентов отбора, варьирующих от 0,05 до 0,80, и для пяти коэффициентов валидности методик предварительного отбора — от нуля (случайный отбор) до 0,50.

Результаты показали впечатляющий прирост продуктивности труда от использования теста при всех этих условиях. Когда отбор на основе теста сравнили со случайным отбором, прирост производительности в долларовом эквиваленте колебался от $97,2 млн при коэффициенте отбора 0,05 до $16,5 млн при коэффициенте отбора 0,80. При валидности предварительного отбора 0,50 соответствующий прирост колебался от 33,3 млн до $5,6 млн. Вероятно, этот прирост можно было бы распространить на ожидаемый срок пребывания в должности вновь нанятых служащих, который для программистов в федеральном правительстве, в среднем, составлял чуть меньше 10 лет. Следует также отметить, что эти оценки основаны на предположении, что отбор начинается с претендентов, имеющих высшие показатели по тесту, и продолжается до тех пор, пока не будет достигнуто заданное значение коэффициента отбора. Иначе говоря, эта процедура предполагает оптимальные условия отбора.

Используя данные переписи населения для определения количества работающих программистами среди населения США, эти исследователи также вычислили оценки эффекта использования данного теста на национальном уровне. В еще более широком

[1] Таблицу, включающую больше значений коэффициентов отбора и валидности, подготовили Нэйлор и Шайн (Naylor & Shine, 1965).

Таблица 6–2

Средние стандартных критериальных показателей принятых на работу в зависимости от валидности теста и коэффициента отбора

Коэффициент отбора	Коэффициент валидности																				
	0,00	0,05	0,10	0,15	0,20	0,25	0,30	0,35	0,40	0,45	0,50	0,55	0,60	0,65	0,70	0,75	0,80	0,85	0,90	0,95	1,00
0,05	0,00	0,10	0,21	0,31	0,42	0,52	0,62	0,73	0,83	0,94	1,04	1,14	1,25	1,35	1,46	1,56	1,66	1,77	1,87	1,98	2,08
0,10	0,00	0,09	0,18	0,26	0,35	0,44	0,53	0,62	0,70	0,79	0,88	0,97	1,05	1,14	1,23	1,32	1,41	1,49	1,58	1,67	1,76
0,15	0,00	0,08	0,15	0,23	0,31	0,39	0,46	0,54	0,62	0,70	0,77	0,85	0,93	1,01	1,08	1,16	1,24	1,32	1,39	1,47	1,55
0,20	0,00	0,07	0,14	0,21	0,28	0,35	0,42	0,49	0,56	0,63	0,70	0,77	0,84	0,91	0,98	1,05	1,12	1,19	1,26	1,33	1,40
0,25	0,00	0,06	0,13	0,19	0,25	0,32	0,38	0,44	0,51	0,57	0,63	0,70	0,76	0,82	0,89	0,95	1,01	1,08	1,14	1,20	1,27
0,30	0,00	0,06	0,12	0,17	0,23	0,29	0,35	0,40	0,46	0,52	0,58	0,64	0,69	0,75	0,81	0,87	0,92	0,98	1,04	1,10	1,16
0,35	0,00	0,05	0,11	0,16	0,21	0,26	0,32	0,37	0,42	0,48	0,53	0,58	0,63	0,69	0,74	0,79	0,84	0,90	0,95	1,00	1,06
0,40	0,00	0,05	0,10	0,15	0,19	0,24	0,29	0,34	0,39	0,44	0,48	0,53	0,58	0,63	0,68	0,73	0,77	0,82	0,87	0,92	0,97
0,45	0,00	0,04	0,09	0,13	0,18	0,22	0,26	0,31	0,35	0,40	0,44	0,48	0,53	0,57	0,62	0,66	0,70	0,75	0,79	0,84	0,88
0,50	0,00	0,04	0,08	0,12	0,16	0,20	0,24	0,28	0,32	0,36	0,40	0,44	0,48	0,52	0,56	0,60	0,64	0,68	0,72	0,76	0,80
0,55	0,00	0,04	0,07	0,11	0,14	0,18	0,22	0,25	0,29	0,32	0,36	0,40	0,43	0,47	0,50	0,54	0,58	0,61	0,65	0,68	0,72
0,60	0,00	0,03	0,06	0,10	0,13	0,16	0,19	0,23	0,26	0,29	0,32	0,35	0,39	0,42	0,45	0,48	0,52	0,55	0,58	0,61	0,64
0,65	0,00	0,03	0,06	0,09	0,11	0,14	0,17	0,20	0,23	0,26	0,28	0,31	0,34	0,37	0,40	0,43	0,46	0,48	0,51	0,54	0,57
0,70	0,00	0,02	0,05	0,07	0,10	0,12	0,15	0,17	0,20	0,22	0,25	0,27	0,30	0,32	0,35	0,37	0,40	0,42	0,45	0,47	0,50
0,75	0,00	0,02	0,04	0,06	0,08	0,11	0,13	0,15	0,17	0,19	0,21	0,23	0,25	0,27	0,30	0,32	0,33	0,36	0,38	0,40	0,42
0,80	0,00	0,02	0,04	0,05	0,07	0,09	0,11	0,12	0,14	0,16	0,18	0,19	0,21	0,22	0,25	0,26	0,28	0,30	0,32	0,33	0,35
0,85	0,00	0,01	0,03	0,04	0,05	0,07	0,08	0,10	0,11	0,12	0,14	0,15	0,16	0,18	0,19	0,20	0,22	0,23	0,25	0,26	0,27
0,90	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09	0,10	0,11	0,12	0,13	0,14	0,15	0,16	0,17	0,18	0,19	0,20
0,95	0,00	0,01	0,01	0,02	0,02	0,03	0,03	0,04	0,04	0,05	0,05	0,06	0,07	0,07	0,08	0,08	0,09	0,09	0,10	0,10	0,11

(Из Brown & Ghiselli, 1953, p. 342.)

исследовании Хантер и Шмидт (Hunter, & Schmidt, 1981) выясняли возможность применения тех же статистических методов для анализа рабочих ресурсов в масштабе страны, с учетом всего спектра профессий. Полученные ими предварительные оценки являются по общему признанию пробными и довольно грубыми, а альтернативные методы, применяемые для такого анализа, в общем дают более низкие оценки (Burke, & Frederick, 1984; U. S. Department of Labor, 1983b; Weekley, Frank, O'Connor, & Peters, 1985). Тем не менее имеющиеся на данный момент результаты убедительно свидетельствуют в пользу того, что эффективные методы распределения людских ресурсов по рабочим местам могут способствовать существенному увеличению валового продукта страны. Природа продуктивного труда, равно как и влияющие на производительность индивидуальные и организационные условия, привлекают все большее внимание исследователей. Прежде всего, это касается развивающейся области исследований критериев, используемых при валидизации тестов, в которой демонстрируются заметные теоретические и методологические достижения (J. P. Campbell, Campbell, & Associates, 1988; Hunter, Schmidt, & Judiesch, 1990; Raju, Burke, & Normand, 1990).

Понятие полезности в теории принятия решений. Именно теория принятия решений позволяет оценить тесты по их эффективности в конкретной ситуации. Такая оценка учитывает не только валидность теста при предсказании определенного критерия, но и ряд других параметров, включая базисную норму и коэффициент отбора. Еще одним важным параметром является относительная *полезность (utility)* ожидаемых результатов, определенным образом оцененная благоприятность или неблагоприятность каждого из них. Отсутствие адекватных методов для приписывания значений результатам с точки зрения единой шкалы полезности служило главным препятствием на пути применения теории принятия решений. В промышленности возможные результаты принимаемых решений часто можно оценить в долларах и центах. Но даже здесь трудно дать денежную оценку некоторым результатам, имеющим непосредственное отношение к доброй воле, социальным отношениям и моральному духу персонала. Решения в области образования должны приниматься с учетом целей учебного заведения, социальных ценностей и других трудно уловимых факторов, а при индивидуальном консультировании — с учетом предпочтений и системы ценностей конкретного человека. Однако уже неоднократно указывалось, что вовсе не теория принятия решений ввела проблему ценностей в процесс принятия решений, она просто сделала ее эксплицитной. Системы ценностей всегда входили составной частью в принимаемые решения, хотя и не сознавались так ясно, да и не согласовывались так систематично, как это имеет место теперь, при использовании теории принятия решений.

Иллюстрацией достижений в развитии методов приписывания ценности альтернативам в моделях принятия решений служит упоминавшееся выше исследование производительности, выполненное Шмидтом, Хантером и их коллегами. Хотя разработанные ими методы предполагают оценку создаваемых работниками товаров и услуг в долларовом эквиваленте, они применимы и для измерения других ценностей. Те же методы, базирующиеся на квантификации человеческих суждений, можно использовать с любой произвольной числовой шкалой, при условии, что эта шкала явно определена и последовательно применяется ко всем результатам. Следует отметить, что требуемые в моделях принятия решений оценки имеют отношение не к абсолютной, а лишь к относительной ценности различных результатов. Всестороннее рассмот-

рение технических аспектов оценки полезности в кадровых решениях можно найти в работе Boudreau (1991).[1]

При выборе стратегии решения цель заключается в максимизации ожидаемой полезности на всем множестве результатов. Схема простой стратегии, представленная на рис. 6–2, поможет прояснить суть метода. На этой схеме изображена стратегия принятия решений в ситуации, отображенной на рис. 6–1, когда в группе претендентов на получение работы проводился всего один тест и на основе сравнения индивидуальных показателей с критическим показателем по этому тесту выносились решения о приеме на работу или отказе. В этой ситуации имеется всего четыре возможных исхода, или результата: правильное/ошибочное принятие и правильное/ошибочное непринятие. Вероятность каждого результата можно вычислить, исходя из числа претендентов, попадающих в каждый квадрант на рис. 6–1. Поскольку в этом примере было всего 100 претендентов, то искомые вероятности, приведенные на рис. 6–2, рассчитываются путем деления каждого из четырех чисел на 100.

Кроме того, нужно знать полезности различных результатов, выраженные в единой шкале. Эти гипотетические величины, полученные с помощью любой оценочной процедуры, приведены в последнем столбце на рис. 6–2. Общую ожидаемую полезность стратегии можно найти, перемножая для каждого из результатов их вероятности и полезности, складывая полученные произведения, а затем вычитая из суммы величину, соответствующую издержкам тестирования. Эта последняя величина высвечивает тот факт, что тесту с низкой валидностью скорее будет отдано предпочтение в ситуации выбора, если он краток, недорог, легко может проводиться малоквалифицированным персоналом и пригоден для группового проведения. Индивидуальному тесту, требующему для своего проведения квалифицированного специалиста или дорогостоящего оборудования, нужно было бы иметь более высокую валидность, чтобы оказаться выбранным для практического использования. В гипотетическом примере на рис. 6–2 величина издержек тестирования, оцененных по шкале полезности, составляет 0,10. Общая ожидаемая полезность (*EU*) этой стратегии вычисляется следующим образом:

$$EU = (0{,}38)(1{,}00) + (0{,}07)(-1{,}00) + (0{,}33)(0) + (0{,}22)(-0{,}50) - 0{,}10 = +0{,}10.$$

Эту *EU* можно затем сравнить с другими *EU*, вычисленными при различных значениях критического показателя, при применении разных тестов (различающихся по валидности и затратам на проведение и обработку данных) или тестовой батареи, а также при использовании различных стратегий принятия решений.[2]

Последовательные стратегии и адаптивный подход. В некоторых ситуациях эффективность теста можно повысить, применяя более сложные стратегии принятия решений, учитывающие большее число параметров. Два примера помогут проиллюстрировать возможности таких стратегий. Во-первых, тесты могут использоваться не только в качестве основания для окончательного решения, но и для *последовательного* принятия решений. В случае простой стратегии (см. рис. 6–1 и 6–2) все решения носят окончательный характер. Напротив, на рис. 6–3 показана двухэтапная последователь-

[1] Что касается других теоретических перспектив оценки полезности, см. Cascio & Morris (1990), Messick (1989, p. 78–81), Sadacca, Campbell, Difazio, Schultz, & White (1990).

[2] Примеры нескольких стратегий принятия решения, в которых показаны все этапы вычислений, можно найти в работе Виггинса (J. S. Wiggins, 1973/1988, chap. 6).

Стратегия	Решение	Результат	Вероятность	Оценка полезности
Провести тест и применить критический показатель	Принять	Правильное принятие	0,38	+1,00
		Ошибочное принятие	0,07	-1,00
	Не принять	Правильное непринятие	0,33	0
		Ошибочное непринятие	0,22	-0,50

Рис. 6–2. Простая стратегия принятия решения

ная стратегия. В качестве теста *A* можно было бы использовать короткий, легкий в проведении, скрининговый тест. На основе результатов этого теста претендентов можно было бы распределить по трем категориям: те, кто будет принят на работу без дополнительных испытаний; те, кто получит окончательный отказ в приеме, и те, кто образует промежуточную группу «сомнительных» случаев. Далее последних можно было бы подвергнуть более интенсивному обследованию с помощью теста *B*, и уже по результатам второго этапа тестирования разделить эту группу на две категории: принятых и не принятых на работу.

Другая стратегия, пригодная для диагностики психологических расстройств, заключается в том, чтобы использовать только две категории, но дополнительно тестировать *всех*, кому на первом этапе тестирования был поставлен положительный диагноз (что указывает на возможную патологию). Эта стратегия уже упоминалась выше в связи с использованием тестов для диагностики патологических состояний при крайне низких базисных нормах.

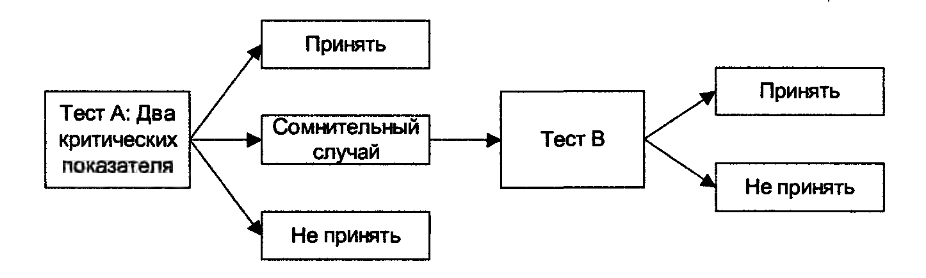


Рис. 6–3. Последовательная стратегия принятия решения

Следует также отметить, что в действительности многие кадровые решения принимаются в соответствии с последовательной стратегией, хотя это и не всегда осознается. Некомпетентные работники, принятые вследствие ошибки прогноза, обычно могут быть уволены по истечении испытательного срока; отчисляются также на ряде этапов не справляющиеся с учебными программами студенты. В таких ситуациях только отрицательное решение оказывается окончательным. Конечно, ошибки отбора, которые затем исправляются, могут дорого обходиться с точки зрения той или иной системы ценностей. Но все-таки они часто сопряжены с меньшими издержками, чем окончательное ошибочное решение.

Вторым условием, влияющим на эффективность психологического теста, является доступность альтернативных методов и возможность *адаптивного подхода*, учитывающего индивидуальные особенности. Примером может служить использование различных программ и методов подготовки персонала в зависимости от уровня их способностей или введение программ коррекции для учеников с определенными трудностями в обучении. В этих условиях стратегия принятия решения в отношении конкретного случая должна строиться с учетом имеющихся сведений о взаимодействии между первоначальным результатом теста и дифференцированным воздействием. Адаптивный подход нередко позволяет значительно повысить процент успешно справляющихся с учебой или работой. Поскольку подбор альтернативного воздействия или режима для конкретного человека является, по существу, проблемой классификации, а не отбора, соответствующая методология будет рассмотрена позже, в разделе, посвященном классификационным решениям.

Приведенные примеры иллюстрируют лишь несколько областей, в которых понятия и принципы теории принятия решений могут помочь в оценке пригодности психологических тестов для специфических целей тестирования. В сущности, эта теория помогла привлечь внимание к сложности комплекса факторов, определяющих выигрыш, который дает использование того или иного теста в конкретной ситуации. Знание коэффициента валидности еще недостаточно для ответа на вопрос, следует ли использовать данный тест, поскольку валидность — лишь один из факторов, подлежащих рассмотрению при оценке влияния теста на эффективность всего процесса выработки решений.[1]

Переменные-модераторы. Валидность теста для определенного критерия может быть разной в подгруппах, различающихся по личным характеристикам входящих в них людей. Классическая психометрическая модель предполагает, что ошибки предсказания являются характеристикой теста, а не тестируемого, и что эти ошибки случайно распределяются между тестируемыми. Гибкость подхода, привнесенная в тестирование теорией принятия решений, побудила к поискам моделей предсказания, учитывающим взаимодействие между тестируемыми и тестами. Такое взаимодействие подразумевает, что один и тот же тест может быть лучшим инструментом предсказания для одних групп или подгрупп, чем для других. Например, данный тест может лучше предсказывать выполнение критериальной деятельности мужчинами, чем женщинами, или кандидатами на рабочие места из более низких, чем из более высо-

[1] Более полное обсуждение следствий теории принятия решений для использования тестов см. в работе Виггинса (J. S. Wiggins, 1973/1988, chap. 6); на техническом уровне эти проблемы обсуждаются в книге Кронбаха и Глесера (Cronbach & Gleser, 1965).

ких социоэкономических уровней. В этих примерах пол и социоэкономический уровень принято называть ***переменными-модераторами (moderator variables)***, так как они ослабляют валидность теста (Saunders, 1956).

Интересы и мотивация могут выполнять функции переменных-модераторов. Так, если кандидатам на рабочие места эта работа мало интересна, они, вероятно, будут выполнять ее без особого усердия, какими бы ни были их показатели по соответствующим тестам способностей. Для таких лиц корреляция между результатами теста способностей и качеством выполнения работы будет низкой, тогда как для заинтересованных и высоко мотивированных такая корреляция может оказаться весьма значительной. Пятидесятые и шестидесятые годы отмечены волной исследований широкого множества переменных, преположительно выполнявших функции модераторов. Серия исследований Гизелли (Ghiselli, 1956, 1960, 1963, 1968) была посвящена прогнозированию уровня выполнения работы. Другие исследователи проверяли гипотезы о роли личностных переменных, особенно в предсказании успешности обучения в колледже (N. Frederiksen, & Gilbert, 1960; N. Frederiksen, & Melville, 1954; Grooms, & Endler, 1960; L. J. Stricker, 1966).

Другая группа относительно устойчивых и согласующихся данных касается влияния половых различий на предсказуемость академической успеваемости. Обзоры, охватывающие несколько сот коэффициентов корреляции из множества источников, говорили о более высокой корреляции между показателями теста способностей и успеваемостью у женщин, чем у мужчин (Gross, Faggen, & McCarthy, 1974; Schmidt, Mellon, & Bylenga, 1978; Seashore, 1962). Эта тенденция была обнаружена как в средних школах, так и в колледжах, хотя в последних она была более выраженной. Имеющиеся данные ничего не говорят о причине таких половых различий в предсказуемости академической успешности, но было бы интересно порассуждать о них в свете других известных половых различий. Если предположить, что ученицы, в общем, оказываются лучше приспосабливающимися и более расположенными к принятию ценностей и норм школьной жизни, их достижения в учебе, вероятно, будут в значительной степени зависеть от их способностей. Если, с другой стороны, предположить, что ученики склонны сосредоточивать свои усилия только на тех занятиях (в школе или вне ее), которые вызывают у них интерес, то эти различия интересов, видимо, будут вносить дополнительную дисперсию в их учебные достижения и тем самым затруднять прогноз успеваемости по результатам теста способностей. Следует, однако, заметить, что половые различия, проявляющиеся в этих коэффициентах валидности, хотя и довольно устойчивы, но, в целом, малы. Кроме того, в более поздних исследованиях обнаруживается тенденция к уменьшению этих различий — результат, который, возможно, отражает меняющиеся аттитюды женщин в конце 1960-х гг. и на протяжении следующего десятилетия.

В общем, ранние ожидания отдачи от изучения переменных-модераторов не оправдались (Abrahams, & Alf, 1972; Pinder, 1973; Zedeck, 1971). Методологический анализ этих исследований обнаруживает множество просчетов. Перекрестная проверка данных на новых выборках часто не подтверждала ранее полученные результаты. Кажется маловероятным, что использование модераторов существенно улучшило бы прогноз, который можно было получить другими средствами. На современном уровне знаний ни одна переменная не может быть признана ослабляющей валидность теста при отсутствии ясных доказательств такого эффекта. Тем не менее понятие переменных-

модераторов может иметь эвристическую ценность для более глубокого понимания индивидуального поведения, например в клинических исследованиях, и для выдвижения свежих гипотез, которые будут проверяться в должным образом контролируемых условиях. Фактически, 1980-е и 1990-е гг. свидетельствовали о возрождении интереса к переменным-модераторам. Некоторые такие переменные, необнаруженные в ранних исследованиях, теперь могут выявляться благодаря применению более совершенных методов статистического анализа данных (Morris, Sherman, & Mansfield, 1986; E. F. Stone, & Hollenbeck, 1989).

Объединение данных различных тестов

Для предсказания практических критериев часто может потребоваться не один, а несколько тестов. Большинство критериев являются комплексными, и их меры зависят от целого ряда различных свойств. Если бы для измерения такого критерия нужно было создать один тест, он получился бы крайне неоднородным. Однако, как уже отмечалось, относительно однородный тест, измеряющий, главным образом, одно свойство, более удовлетворителен, так как дает более однозначные результаты (глава 5). Поэтому обычно предпочтительней пользоваться серией из нескольких относительно однородных тестов, каждый из которых нацелен на какой-то один аспект критерия, чем одним тестом, представляющим собой мешанину самых разнородных заданий.

Когда несколько специально подобранных тестов применяются вместе для предсказания одного-единственного критерия, такую совокупность тестов называют *тестовой батареей (test battery)*. Главная проблема, возникающая при использовании таких батарей, касается способа объединения показателей по отдельным тестам при выработке решения в каждом индивидуальном случае. Для этой цели обращаются к двум основным видам процедур, а именно использованию уравнения множественной регрессии и анализу профиля (*profile analysis*). Когда тесты применяются в интенсивном исследовании индивидуальных случаев, например при уточнении клинического диагноза, консультировании или при оценке руководителей высшего звена, проводящий тестирование специалист по большей части пользуется показателями теста, не прибегая к их статистическому анализу. Составляя заключение или давая рекомендации, он интерпретирует конкретный набор показателей и объединяет результаты отдельных тестов, опираясь на свою проницательность, прошлый опыт и теоретические соображения.

Уравнение множественной регрессии. Уравнение множественной регрессии позволяет получить числовую оценку прогнозируемого критерия для каждого испытуемого на основе его показателей по всем тестам батареи. Следующее уравнение регрессии иллюстрирует применение этой процедуры для предсказания успеваемости старшеклассника по математическим дисциплинам на основе его показателей по вербальному (V), числовому (N) и логическому (R) тестам:

$$\text{Успехи в математике} = 0{,}21V + 0{,}21N + 0{,}32R + 1{,}35.$$

В этом примере тестовые показатели и оценка критерия выражаются в станайнах, но для этой цели можно было бы использовать любую другую шкалу показателей.

В приведенном выше уравнении выраженный в станайнах показатель ученика по каждому из трех тестов умножается на соответствующие веса, заданные в этом уравнении. Сумма трех произведений плюс константа (1,35) дает прогнозируемое положение ученика (в шкале станайнов) по математике.

Предположим, Бетти Джонс получила следующие показатели в станайнах:

Вербальный тест: 6
Числовой тест: 4
Логический тест: 8

Ожидаемые успехи по математике у этой ученицы определяются следующим образом:

Успехи в математике = (0,21) (6) + (0,21) (4) + (0,32) (8) + 1,35 = 6,01.

Итак, прогнозируемый станайн Бетти примерно равен 6. Напомним (глава 3), что станайн 5 соответствует среднему уровню выполнения критериальной деятельности. Значит, Бетти, вероятно, будет иметь по математике оценки несколько выше среднего. Ее очень высокий результат по логическому тесту ($R = 8$) и превышающий средний уровень результат по вербальному тесту ($V = 6$) компенсируют невысокую скорость и точность вычислений ($N = 4$).

Конкретные вычислительные процедуры применительно к уравнениям регрессии можно найти в учебниках по статистике для психологов (например, D. C. Howell, 1997; Runyon, & Haber, 1991). По существу, такое уравнение основано на корреляции каждого теста с критерием и корреляциях тестов между собой. Очевидно, что тесты, сильнее коррелирующие с критерием, должны получить больший вес. Столь же важно, однако, учитывать корреляцию каждого теста с другими тестами батареи. Высокая корреляция указывает на ненужное дублирование одного теста другим, ибо это означает, что тесты в значительной мере направлены на один и тот же аспект критерия. Включение двух таких тестов не повышает существенно валидности всей батареи, даже если оба они тесно коррелируют с критерием. В этом случае один из этих тестов столь же эффективен, как и пара, поэтому в батарее следует оставить только один тест.

Однако даже после того, как случаи наиболее выраженного дублирования тестов в батарее устраняются, оставшиеся тесты все равно будут в той или иной степени коррелировать друг с другом. Для максимизации прогнозирующей силы тесты, вносящие более «уникальный» вклад в полную батарею, должны получать больший вес по сравнению с тестами, частично дублирующими функции других тестов батареи. При расчете коэффициентов уравнения множественной регрессии каждый тест получает вес, прямо пропорциональный его корреляции с критерием и обратно пропорциональный корреляции с другими тестами. Это значит, что максимальный вес получит тест, обладающий наибольшей валидностью и в наименьшей степени дублирующий остальную часть батареи.

Валидность полной батареи можно найти путем вычисления коэффициента множественной корреляции (R) между входящими в нее тестами и критерием. Этот вид корреляции дает оценку максимальной предсказуемостной эффективности, которой можно добиться от данной тестовой батареи при условии, что каждый входящий в нее тест получает оптимальный — с точки зрения предсказания критерия — вес. Оптимальные веса как раз и определяются по уравнению регрессии.

Следует иметь в виду, что эти веса являются оптимальными только для конкретной выборки, по результатам обследования которой они были найдены. Поскольку в используемых при определении весов коэффициентах корреляции всегда присутствуют случайные (несистематические) ошибки, весовые коэффициенты регрессии могут меняться от выборки к выборке. Поэтому батарею следует подвергнуть перекрестной валидизации, коррелируя прогнозируемые показатели критерия с его фактическими показателями в новой выборке. Для оценки степени *естественной убыли (shrinkage)* множественной корреляции, которой можно ожидать при применении уравнения регрессии к другой выборке, существуют специальные формулы, но, если есть возможность, предпочтительней провести эмпирическую проверку. В целом же, чем больше выборка, по которой определялись веса, тем меньшей будет эта естественная убыль.[1]

В определенных ситуациях прогностическую валидность батареи можно повысить, включая в уравнение регрессии переменную, которая представляет тест, имеющий нулевую корреляцию с критерием и высокую корреляцию с одним из тестов батареи. Такая необычная ситуация возникает, когда тест, не коррелирующий с критерием, действует как *переменная-подавитель (suppressor variable)*, устраняющая или гасящая нерелевантную дисперсию показателей коррелирующего с ним теста. Например, понимание читаемого текста может тесно коррелировать с показателями теста математических или механических способностей, так как выполнение заданий этих тестов требует понимания сложных письменных инструкций. Даже если понимание текста не имеет отношения к прогнозируемой трудовой деятельности, оно, будучи необходимым для выполнения тестов, вносит дисперсию ошибок в результаты и снижает прогностическую валидность этих тестов. Проведя тест на понимание читаемого и включив его показатели в уравнение регрессии, мы устраним эту дисперсию ошибок и повысим валидность батареи. Переменная-подавитель входит в уравнение регрессии с отрицательным знаком. Поэтому чем выше показатель конкретного человека по тесту понимания читаемого текста, тем большая величина вычитается из его показателя по тесту математических или механических способностей. Однако в любой ситуации для исключения нерелевантной дисперсии предпочтительней использовать более прямую процедуру пересмотра теста, чем косвенный способ статистического устранения такой дисперсии с помощью переменной-подавителя. И только в тех случаях, когда внесение изменений в тест невозможно или недопустимо, следует рассмотреть вариант использования переменных-подавителей. В таких случаях эффект переменной-подавителя нужно всегда проверять на новой выборке.

Анализ профиля и критические показатели. В дополнение к анализу индивидуальных профилей, применяемому в клиническом обследовании, паттерн, или конфигурацию тестовых показателей, полученных с помощью батареи для отбора персонала, можно оценивать на основе множественного критерия, представленного набором критических показателей. Если коротко, то этот способ заключается в установлении минимального критического показателя по каждому тесту батареи. Когда применяется

---

[1] При определенных условиях в качестве весовых коэффициентов регрессии предпочтительней использовать «удельные веса» или другие альтернативы. Краткий обзор исследований различных схем взвешивания см. в Dunnette & Borman (1979).

строгий вариант этого метода, всякий, кто не достигает такого минимального уровня *хотя бы по одному* из тестов, считается не прошедшим тестирования. При выборе тестов и установлении критических показателей, подходящих для определенной профессии, обычно исходят не только из величины коэффициентов валидности тестов. Если бы в расчет принимались только тесты со значимыми коэффициентами валидности, то могли оказаться неучтенными существенные навыки или способности, которыми должны обладать все представители определенной профессии. Поэтому необходимо рассматривать и те способности, которые должны быть хорошо развиты у тестируемых как единой профессиональной группы, даже если индивидуальные различия между ними, наблюдающиеся выше критериального минимума, никак не связаны с успешностью работы. Кроме того, представители некоторых профессий могут представлять собой настолько однородную группу по ключевой переменной, что диапазон индивидуальных различий оказывается слишком узким, чтобы обеспечить значимую корреляцию между показателями теста и критерием.

Наиболее полной иллюстрацией применения метода множественных критических показателей может служить Батарея тестов общих способностей (*GATB*)[1], разработанная Службой занятости США для целей профконсультирования и профпросвещения в ее региональных отделах (U. S. Department of Labor, 1970). Девять показателей способностей, которые дает эта батарея и которые рассматриваются применительно к каждой профессии, были выбраны на основе корреляции с критерием, среднего и стандартного отклонения показателей представителей конкретных профессий, а также качественных оценок специалиста по анализу трудовых операций.

Наиболее сильный аргумент в пользу применения множественных критических показателей, а не уравнения регрессии, основывается на возможности существования компенсирующих показателей (*compensatory scores*). Другими словами, серьезная недостаточность в одном навыке может остаться незамеченной в суммарном показателе индивидуума по тестовой батарее вследствие высокого показателя по другому тесту. Если эта недостаточность относится к навыку, который является решающим для выполнения определенной работы, отобранный кандидат потерпит неудачу, независимо от его способностей в других областях. Однако такой ситуации можно избежать, установив один или несколько критических навыков, необходимых в определенной профессии, и применяя критический показатель только в соответствующих тестах. В большинстве же тестов обычно предпочтительнее сохранять актуальный, фактический показатель, поскольку чем выше тестовый показатель конкретного человека, тем выше, в общем, будет эффективность его работы. Для большинства профессий связь между прогнозирующим показателем и критериальной деятельностью носит линейный характер. Следует добавить, что именно широкие исследования с использованием батареи *GATB* снабдили нас надежными данными о линейности такой связи (Coward, & Sackett, 1990; Hartigan, & Wigdor, 1989; Hawk, 1970). При этих условиях отбор персонала на основе фактической величины тестовых показателей обеспечивает более высокую эффективность работы, чем отбор на основе превышения минимальных критических показателей.

---

[1] Эта широко используемая тестовая батарея рассматривается в главе 17, в связи с применением тестов в сфере промышленности и управления.

Использование тестов для принятия классификационных решений

Характер классификации. Психологические тесты могут использоваться для целей отбора, расстановки и распределения (или классификации). При *отборе (selection)* каждый индивид либо принимается, либо не принимается. Решения о зачислении абитуриента в колледж, принятии кандидата на работу или направлении новобранца на офицерские курсы — все это примеры отбора. Когда отбор производится в несколько этапов, его начальные стадии часто называют *отсеиванием*, или *скринингом*, а термин «отбор» сохраняют за более интенсивными заключительными стадиями. Термин «отсеивание» может к тому же употребляться для обозначения любой формы быстрого и приблизительного отбора, даже если он не сопровождается углубленными процедурами отбора.

Расстановка и распределение отличаются от отбора тем, что их осуществление не связано с выбыванием кого бы то ни было из участников программы. Для всех участников определяются соответствующие места или «комбинации условий» с тем, чтобы максимизировать конечный результат. В случае *расстановки (placement)* назначения могут основываться на единственном показателе, который можно получить с помощью одного теста — скажем, математического теста достижений. Если применяется батарея тестов, ту же роль может сыграть совокупный показатель, вычисленный по уравнению регрессии. Примерами расстановки могут служить решения о делении первокурсников по данным теста достижений на группы для изучения математики, назначении канцелярских работников на требующие разного уровня компетентности и ответственности должности или определении степени тяжести психически больных в целях назначения соответствующей терапии. Очевидно, что в каждом из этих решений применяется лишь один критерий и что определение места конкретного человека определяется его положением на одной-единственной шкале прогнозирующего показателя.

В отличие от расстановки, при *распределении (classification)* во внимание принимается два критерия или более. Так, в армии распределение — одна из главных проблем, поскольку каждый новобранец должен быть приписан к той военной специальности, где его служба будет наиболее эффективной. Решения о распределении людских ресурсов столь же необходимы в промышленности, когда вновь нанятые сотрудники направляются на курсы подготовки для последующего выполнения разного рода работ. Еще одним примером может служить консультирование студентов по вопросу выбора программы обучения (естественные науки, гуманитарные науки, и т. д.) или области специализации. Консультирование основывается в значительной степени на распределении, так как клиенту сообщаются его шансы на успех в разных академических программах или профессиях. Клинический диагноз также представляет собой проблему распределения, ибо в этом случае главной целью каждого диагноза является решение о наиболее подходящем курсе лечения.

Если расстановка может осуществляться на основе одного или нескольких прогнозирующих показателей, то распределение требует множественных предикторов, валидность которых устанавливается отдельно по каждому критерию. Классификационная батарея требует разных уравнений регрессии для каждого критерия. Одни тесты могут быть представлены во всех уравнениях, хотя и с разными весами, другие —

только в одном или двух, а в остальных уравнениях их веса равны или близки к нулю. Таким образом, используемая комбинация тестов из состава батареи и их веса меняются в зависимости от критерия. Один из ранних образцов такой классификационной батареи является тестовая батарея, разработанная в военно-воздушных силах США для распределения личного состава по различным курсам специальной подготовки. Эта батарея, состоящая как из тестов типа «бумага—карандаш», так и из аппаратных тестов, обеспечивала получение выраженных в станайнах показателей для пилотов, штурманов, бомбардиров и ряда других специалистов ВВС. Находя ожидаемые значения критериальных показателей по различным уравнениям регрессии, можно было предсказать, например, что данного человека лучше готовить к специальности пилота, чем штурмана. Современный образец гораздо более широкой батареи — Проект *А*, или Проект отбора и распределения специалистов сухопутных войск США (J. P. Campbell, 1990b).

Дифференциальная валидность. При оценивании классификационной батареи большое значение придается ее дифференциальной валидности по отдельным критериям. Цель такой батареи — предсказание *разницы* в выполнении каждым человеком двух или более видов профессиональной деятельности, учебных программ или в других критериальных ситуациях. Тесты, из которых составляются такие классификационные батареи, должны давать сильно различающиеся коэффициенты валидности для отдельных критериев. Например, применительно к задаче классификации по двум критериям идеальный тест имел бы высокую корреляцию с одним критерием и нулевую (или, еще лучше, отрицательную) — с другим. Тесты общего интеллекта сравнительно мало пригодны для целей распределения, так как они примерно одинаково прогнозируют успех в большинстве областей деятельности. Поэтому их корреляции с подлежащими дифференциации критериями были бы слишком сходными. Человека, набравшего высокий балл по такому тесту, пришлось бы классифицировать как подходящего для любого назначения, и было бы невозможно предсказать, где он преуспеет больше. В классификационной батарее должно быть несколько тестов, являющихся хорошими предикторами критерия *А* и плохими предикторами критерия *В*, и несколько других тестов — плохих предикторов критерия *А*, но зато хороших предикторов критерия *В*.

Для отбора тестов с целью максимизации дифференциальной валидности классификационной батареи разработаны специальные статистические методы (Brogden, 1946а, 1951, 1954; Horst, 1954; Mollenkopf, 1950b; Zeidner, & Johnson, 1991). Однако когда число критериев больше двух, проблема сильно усложняется, и для таких случаев нет чисто аналитического решения. На практике применяют различные эмпирические методы, чтобы приблизиться к желаемым целям. Исчерпывающий анализ сложностей, связанных с решением задачи классификации, дан Кэмпбеллом (J. P. Campbell, 1990а, pp. 715–721).

Множественные дискриминантные функции. Альтернативный подход к проблеме принятия классификационных решений основан на применении множественной дискриминантной (или классифицирующей) функции (French, 1966). По существу, это математический метод для определения того, насколько показатели конкретного человека по всему набору тестов приближаются к показателям, типичным для представителей определенной профессии, учебной программы, психиатрического синдро-

ма или другой категории. После чего этого человека можно было бы отнести к той специфической группе, с которой он обладает наибольшим сходством. Если уравнение регрессии позволяет предсказать степень успеха в каждой области, то метод множественной дискриминантной функции позволяет рассматривать всех тестируемых в рамках одной категории как обладающих равным статусом. Групповое членство — единственные критериальные данные, используемые этим методом. Классифицирующая функция полезна в тех случаях, когда критериальные показатели недоступны и можно установить только групповую принадлежность. Валидизация некоторых тестов, например, производится путем проведения их с людьми, занятыми в разных профессиях, хотя при этом отсутствуют какие-либо меры успешности работы для конкретных людей в каждой такой профессиональной области.

Дискриминантную функцию целесообразно применять и тогда, когда связь между критерием и одним или несколькими предикторами носит нелинейный характер. Например, для некоторых черт личности может существовать оптимальный диапазон, отвечающий данной профессии. Лица с большей или меньшей выраженностью такой черты оказались бы, таким образом, в невыгодном положении. Разумно ожидать, что, скажем, продавцы с умеренно высоким уровнем социального доминирования скорее всего будут преуспевать в работе и что их шансы на успех будут снижаться по мере отклонения их тестовых показателей в любую сторону от этой оптимальной области. С помощью дискриминантных функций мы, в общем, и отбираем тех, чьи показатели попадают в границы оптимальной области, тогда как использование уравнения регрессии заставило бы нас ожидать наилучшей работы от продавцов с максимальным показателем социальной доминантности[1]. Разумеется, при отрицательной корреляции между прогнозирующим показателем и критерием уравнение регрессии дало бы более благоприятный прогноз для лиц с низкими тестовыми показателями. Но все равно в этом случае нет прямого способа получить максимальную оценку для промежуточного значения тестового показателя. Хотя во многих случаях оба этих метода дают одинаковые результаты, существуют ситуации, когда одни и те же лица могут оказаться отнесенными к разным категориям при их распределении на основе уравнения регрессии и дискриминантной функции. Для большинства целей психологического тестирования применение уравнения регрессии более эффективно, однако при некоторых обстоятельствах дискриминантная функция лучше подходит для получения необходимой информации.

Максимизация использования талантов. Дифференциальное прогнозирование критериев с помощью батареи тестов позволяет полнее использовать людские ресурсы, чем при применении одного общего теста или совокупного показателя, вычисляемого по уравнению регрессии. Как видно из таблиц Тейлора—Расселла и из других примеров данной главы, эффективность любого теста при отборе персонала для выполнения определенной работы зависит от коэффициента отбора. При принятии классификационных решений мы работаем с меньшими величинами коэффициента отбора и, следовательно, имеем возможность назначить на каждую должность более квалифицированных людей. Если из 100 претендентов предполагается принять по 10 человек на каждую из двух должностей или специальностей, то при использовании отдельных предикторов для каждой из них коэффициент отбора составит 10 %. Если

[1] Это утверждение авторов справедливо только в отношении *линейной* регрессии. — *Примеч. науч. ред.*

же используется единственный предиктор (такой, как тест общего интеллекта), то коэффициент отбора составит уже 20 %, поскольку нам ничего не остается, как взять на работу 20 человек с наибольшими показателями.

Даже когда предикторы обеих специальностей тесно коррелируют между собой, так что некоторые из претендентов могли быть приняты как на одну, так и на другую работу, использование отдельных предикторов все равно дает значительный выигрыш. Эта ситуация проиллюстрирована в табл. 6–3, где приведены средние стандартные критериальные показатели работников, принятых на каждую из двух должностей при использовании стратегии отбора (единственный предиктор) и стратегии распределения (или, иначе говоря, классификации) с двумя различными предикторами, валидность каждого из которых определена относительно собственного профессионального критерия. Если бы работников принимали наугад, без всякого отбора, средний стандартный показатель в этой шкале был бы равен нулю. Аналогичный результат получился бы и в том случае, если бы коэффициент отбора на каждую должность составлял 50 %, так что всех 100 % подавших заявление пришлось бы принять на работу. Заметим, что даже в этих условиях, как видно из нижней строки таблицы, использование двух предикторов привело бы к повышению среднего уровня выполнения работы. При двух некоррелирующих предикторах оценка этого уровня была бы равна 0,31 (т. е. почти на 1/3 стандартного отклонения выше среднего уровня выполнения работы теми, кого приняли наугад). С ростом корреляции между предикторами эффективность работы отобранных на их основе лиц снижается, но все еще остается выше эффективности случайно набранных работников даже при корреляции 0,80. При более низких значениях коэффициента отбора, разумеется, можно набрать более квалифицированный персонал. Однако, как видно из табл. 6–3, средний уровень выполнения работы при любом значении коэффициента отбора остается выше для принятых на работу при использовании стратегии распределения, чем стратегии отбора.

Таблица 6–3

Средние стандартные критериальные показатели лиц, назначенных на каждую из двух должностей при использовании стратегий отбора или распределения

Коэффициент отбора на каждую должность (%)	Отбор: один предиктор	Распределение (классификация): два предиктора с коэффициентами взаимокорреляции, равными:				
		0	**0,20**	**0,40**	**0,60**	**0,80**
5	0,88	1,03	1,02	1,01	1,00	0,96
10	0,70	0,87	0,86	0,84	0,82	0,79
20	0,48	0,68	0,67	0,65	0,62	0,59
30	0,32	0,55	0,53	0,50	0,46	0,43
40	0,18	0,42	0,41	0,37	0,34	0,29
50	0,00	0,31	0,28	0,25	0,22	0,17

(Перепечатано в сокращении из Brogden, 1951, p. 182)

Практической иллюстрацией преимуществ стратегий распределения служит использование показателей Областей пригодности (*Aptitude Areas*) при распределении личного состава по военным специальностям в сухопутных войсках США (Maier, & Fuchs, 1973). В этом исследовании каждая Область пригодности соответствовала группе армейских профессий, требующих сходного паттерна способностей, знаний и интересов. Для определения показателя военнослужащего в каждой Области пригод-

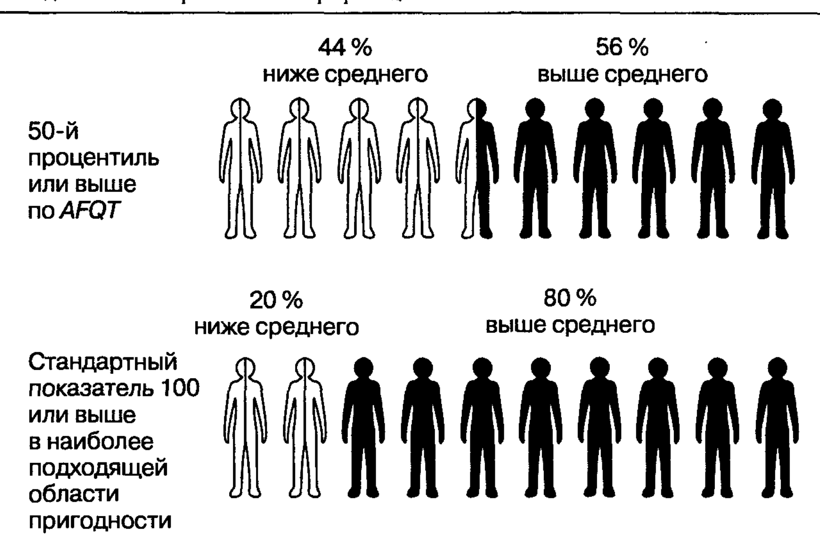


Рис. 6–4. Процент получивших показатели выше среднего в тесте *AFQT* и в наиболее подходящих Областях пригодности по Армейской классификационной батарее в выборке 7500 добровольно поступающих на военную службу

(По данным U. S. Army Research Institute for the Behavioral and Social Sciences. С любезного разрешения J. E. Uhlaner.)

ности использовалось от трех до пяти тестов из 13-тестовой классификационной батареи. На рис. 6–4 представлены результаты исследования 7500 добровольно поступающих на военную службу, в котором сравнивалась эффективность использования показателей Области пригодности и общего теста отсеивания, так называемого Квалификационного теста вооруженных сил (*Armed Forces Qualification Test* [*AFQT*]). Отметим, что только 56 % этой группы достигли или превысили 50-й процентиль по *AFQT*, в то время как 80 % достигли или превысили средний стандартный показатель, равный 100, в своей наилучшей Области пригодности. Таким образом, когда людей распределяют по конкретным рабочим местам на основе необходимых для выполнения такой работы способностей, подавляющее большинство способно справиться с ней на уровне не хуже или даже лучше среднего для всей выборки. Казалось бы, невозможно почти каждому быть выше среднего, но это достигается благодаря тому, что почти каждый превосходит средний уровень хотя бы *в какой-то одной* способности.

По сути то же самое было наглядно показано при изучении совершенно иной совокупности — одаренных детей (Feldman, & Bratton, 1972). В демонстрационных целях 49 детей из двух обычных 5-х классов оценили по 19 показателям, до этого использовавшимися при отборе учеников для специальных программ работы с одаренными детьми. Среди этих показателей были общие показатели группового теста интеллекта и батареи учебных достижений, оценки по тестам отдельных способностей и учебных навыков, скажем чтения и арифметики, показатели теста творческого мышления, оценки по музыке и рисованию, а также результаты выбора учителями наиболее «одаренных» и «творческих» детях в классе. Когда по каждому критерию было выделено по пять лучших учеников, вместе они составили 92 % группы. Тем самым еще раз было показано, что применение многомерных критериев позволяет установить превосходство в каких-то областях почти всех членов группы.

Статистический анализ систематической ошибки теста

Проблема. Если мы хотим использовать тесты для прогнозирования результатов в каких-то будущих ситуациях, скажем для предсказания академической успеваемости абитуриента или успешности работы кандидата на определенную должность, нам нужны тесты с высокой прогностической валидностью относительно специфического критерия. Это требование обычно упускают из вида при разработке так называемых культурно-свободных тестов (обсуждаемых далее в главах 9 и 12). Стремясь включить в такие тесты только функции, общие для разных культур или субкультур, мы можем отобрать содержание, которое имеет мало отношения к какому-либо из прогнозируемых критериев. Лучшим решением было бы подобрать релевантное критерию содержание, а затем исследовать возможные популяционные различия в эффективности теста относительно намеченной цели. Коэффициенты валидности, весовые коэффициенты регрессии и критические показатели могут меняться в зависимости от биографических данных тестируемых. Эти величины следует поэтому проверять в подгруппах, для которых есть основание ожидать влияния таких данных. Такого рода возможные различия между подгруппами можно было бы признать особым случаем роли переменных-модераторов, обсуждавшихся в предыдущем разделе. И следует напомнить, что поиск значимых и устойчивых эффектов модераторов дал неутешительные результаты. В данном разделе мы рассмотрим конкретные приложения этого вида анализа к различным группам меньшинств в США.

Заметим, однако, что прогностические характеристики тестовых показателей меньше зависят от различий в культурах, если тест внутренне связан с критериальной деятельностью. Если вербальный тест используется для прогноза невербальной профессиональной деятельности, он может случайно оказаться валидным в одной культурной группе вследствие традиционных ассоциаций прошлого опыта работы в такой культуре. Между тем в группе с иными культурными традициями этот тест может полностью потерять свою валидность. С другой стороны, тест, который выборочно проверяет само критериальное поведение или измеряет необходимые для работы навыки, вероятно, будет сохранять свою валидность в различных группах.

Начиная с середины 1960-х гг. происходит быстрое накопление данных исследований, посвященных возможным этническим различиям в прогностическом значении тестовых показателей.[1] Подавляющее большинство исследований, проведенных на сегодняшний день, касались афроамериканцев, и лишь в некоторых из них затрагивались другие этнические меньшинства. Изучавшиеся проблемы обычно объединяются под общей рубрикой: *систематическая ошибка теста (test bias)*. В данном контексте термин «систематическая ошибка» употребляется в твердо установившемся статистическом смысле, для обозначения постоянной, или систематической, ошибки в противоположность случайной ошибке. Тот же самый смысл мы вкладываем в выражение смещенная (т. е. необъективная, пристрастная) выборка, противополагая ее случайной выборке. Главные вопросы, поставленные в связи с систематической ошибкой

[1] Из всей этой обширной литературы можно упомянуть лишь несколько репрезентативных исследований. В том, что касается общей характеристики данной проблемы и анализа ее многочисленных аспектов, мы рекомендуем следующие работы: N. S. Cole & Moss (1989), Hunter, Schmidt, & Rauschenberger (1977), C. R. Reynolds & Brown (1984).

теста, имеют отношение к коэффициенту валидности (систематическая ошибка наклона) и к соотношению между групповыми средними по тесту и по критерию (систематическая ошибка интерцепта). Эти вопросы будут рассмотрены в двух следующих разделах.

Систематическая ошибка наклона. Чтобы облегчить понимание технических аспектов систематической ошибки теста, начнем с диаграммы рассеяния, или двумерного распределения (см. главу 4, особенно рис. 4–3). Правда, в данном случае по горизонтальной оси *(X)* откладываются тестовые показатели, а по вертикальной *(Y)* — критериальные показатели, такие как средняя успеваемость в колледже или индекс производительности труда. Напомним, что «палочки», изображающие положение каждого индивидуума относительно теста и критерия, в своей совокупности показывают направление и общую величину корреляции между этими двумя переменными. Линия наилучшего согласия, проведенная через множество кодировочных «палочек», называется линией регрессии, а ее уравнение — уравнением регрессии. В этом примере уравнение регрессии содержит только один прогнозирующий показатель. Уравнения множественной регрессии, о которых говорилось выше, содержат несколько прогнозирующих показателей, но принцип остается тем же самым.

Когда и тестовые, и критериальные показатели выражены в виде стандартных показателей (*SD* = 1,00), *угловой коэффициент* (или попросту — *«наклон»*) линии регрессии равен коэффициенту корреляции. По этой причине, когда тест дает значимо различающиеся коэффициенты валидности в двух группах, это различие называют систематической ошибкой наклона. Этот вид групповых различий часто описывают как «дифференциальную валидность». Некоторые исследователи используют также термин «одно-групповая валидность» (*single-group validity*) по отношению к тесту, коэффициент валидности которого достигает статистической значимости в одной группе, но оказывается незначимым в другой.

На рис. 6–5 дается схематическое изображение линий регрессии для нескольких двумерных распределений.[1] Эллипсами обозначены области, в границах которых сосредоточены закодированные «палочками» представители каждой выборки. Случай 1 соответствует двумерным распределениям двух групп с различными средними прогнозирующего (тестового) показателя, но с идентичными линиями регрессии между предиктором (тестом) и критерием. В данном случае тест не дает систематической ошибки, так как любой данный тестовый показатель (*X*) соответствует одинаковому критериальному показателю в обеих группах. Случай 2 иллюстрирует систематическую ошибку наклона, с более низким коэффициентом валидности для группы меньшинства.

В исследованиях дифференциальной валидности общей помехой часто оказывается значительно меньшее количество испытуемых в выборке меньшинства, чем в

[1] Показанный на рис. 6–5 тип анализа систематической ошибки получил название «модель Клири», поскольку был применен Клири (Cleary, 1968) в широко цитируемом исследовании показателей Теста академических способностей Совета колледжей у студентов из различных меньшинств. Подходящие математические процедуры разработали Галликсен и Уилкс (Gulliksen & Wilks, 1950), а Хамфрис (Humphreys, 1952) предложил применить их для сравнения групп, различающихся по этнической принадлежности и полу. Диаграммы на рис. 6–5 взяты (с некоторыми упрощениями) из исследования М. Гордона (M. A. Gordon, 1953), проведенного под руководством Хамфриса в военно-воздушных силах США.

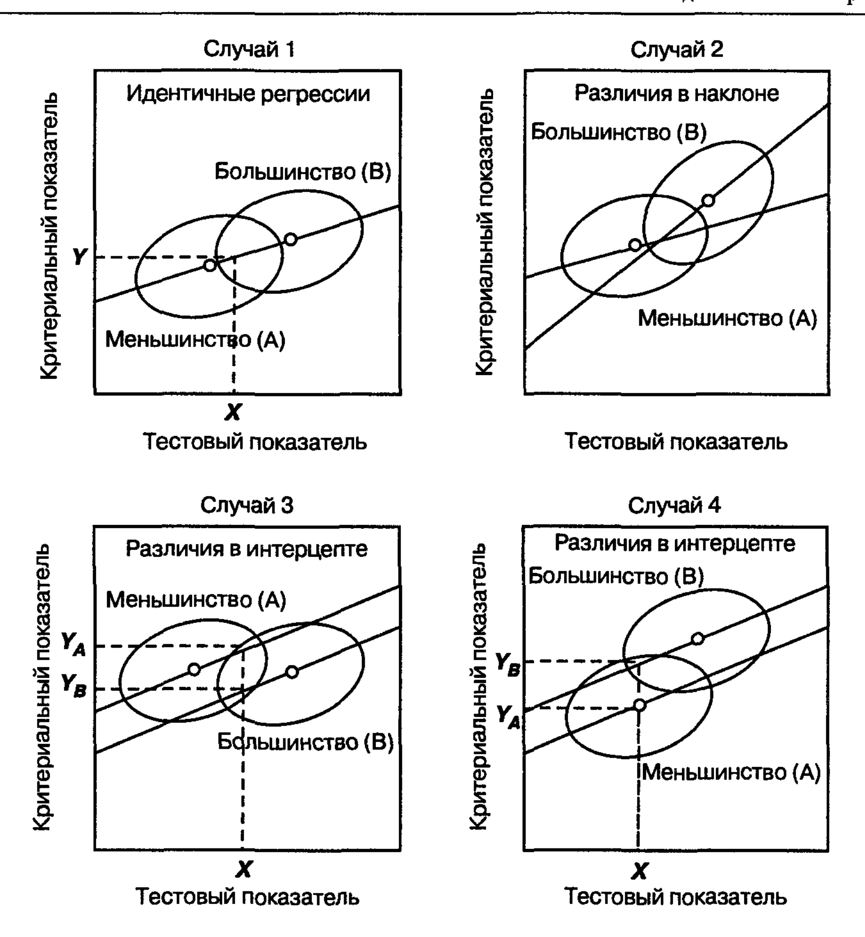


Рис. 6–5. Систематические ошибки наклона и интерцепта при прогнозировании критериальных показателей. Эллипсами выделены области, в которые попадают члены каждой группы при построении диаграммы рассеяния тестовых показателей относительно результатов критериальной деятельности.
(Случаи 1, 2 и 4 взяты — с некоторыми изменениями — из работы М. А. Gordon, 1953, р. 3)

выборке большинства. При этих условиях один и тот же коэффициент валидности может оказаться статистически значимым в выборке большинства и незначимым в выборке меньшинства (так называемая одно-групповая валидность). При выборке в 100 человек, например, коэффициент корреляции 0,27 значим на уровне 0,01, тогда как при 30 испытуемых тот же коэффициент далек от минимальной величины, необходимой для достижения значимости даже на уровне 0,05. По этой причине в исследованиях дифференциальной валидности рекомендуется определять не статистическую значимость коэффициентов валидности раздельно для каждой группы, а оценивать статистическую значимость различий между такими коэффициентами (Humphreys, 1973). В противном случае можно было бы легко «доказать», что тест валиден, скажем, для нас, белых, и не валиден для черных. Все, что для этого потребовалось бы, — это достаточно большая группа белых и относительно небольшая группа черных!

Более тонкий статистический анализ результатов 19 опубликованных исследований, в которых сообщаются коэффициенты валидности для выборок работающего белого и черного населения США, подвергает серьезному сомнению выводы некоторых более ранних исследований (Schmidt, Berner, & Hunter, 1973). Учитывая найденные величины валидности и объемы выборок в каждом исследовании, удалось наглядно доказать, что различия коэффициентов валидности, обнаруженные между выборками черного и белого населения, не отличаются от случайных. Этот вывод был подтвержден результатами последующего, более широкого анализа, охватывающего 39 исследований (Hunter, Schmidt, & Hunter, 1979). Тема различающейся валидности тестов для претендентов на получение работы из основной группы населения и меньшинств вызывала непрекращающиеся дискуссии на протяжении более десятка лет. Некоторые исследователи отмечали, что полученные здесь результаты, из-за методологических недостатков, просто не позволяли делать каких-то определенных выводов. Примечательно, однако, что в хорошо спланированных, крупномасштабных исследованиях на выборках работников промышленности (J. T. Campbell, Crooks, Mahoney, & Rock, 1973) и личного состава вооруженных сил (Maier, & Fuchs, 1973) никаких данных в подтверждение дифференциальной валидности получено не было. В общем, чем совершеннее исследование в методологическом отношении, тем менее вероятно обнаружить в нем дифференциальную валидность.

Сходные результаты были получены в многочисленных исследованиях черных и белых студентов колледжей (Breland, 1979). Коэффициенты валидности проводимого Советом колледжей Теста академических способностей и других тестов, по результатам которых осуществляется прием в колледж, обычно столь же высоки для черных, как и для белых абитуриентов, а иногда и выше. Это соотношение обнаруживается при анализе выборок черных и белых студентов, обучающихся как в одних колледжах, так и раздельно. Изучая совершенно иной уровень образования, Митчелл (B. C. Mitchell, 1967) исследовал валидность двух тестов готовности к школьному обучению относительно показателей первоклассников по тесту достижений на конец учебного года. В больших выборках протестированных черных и белых детей валидность как общих показателей, так и показателей по субтестам оказалась почти одинаковой для этих двух этнических групп, несмотря на обнаружившуюся тенденцию быть несколько выше у черных детей. Если обобщить сказанное, то исчерпывающие научные обзоры и критический анализ опубликованных исследований не дали оснований для поддержки гипотезы о том, что тесты способностей менее валидны для черных, чем для белых при прогнозировании результатов учебной или профессиональной деятельности (Hunter, Schmidt, & Rauschenberger, 1984; Linn, 1978).

Хотя сопоставимых исследований, проведенных с другими меньшинствами, значительно меньше, сходные результаты были получены для испаноязычных американцев применительно как к образовательному тестированию, так и к тестированию при приеме на работу (Breland, 1979; Duran, 1983, 1989; Pennock-Román, 1990; Schmidt, Pearlman, & Hunter, 1980). Однако в отношении испаноязычных американцев интерпретация тестовых показателей осложняется варьированием степени двуязычия и влиянием социокультурных (связанных с исторической ро́диной) переменных; и то и другое сказывается не только на выполнении тестов, но и на академических и профессиональных достижениях. При этих условиях вряд ли можно надеяться, что все это не повлияет на прогностическую валидность. В четко спланированном обзоре опубликованных исследований использования тестов при приеме в колледж, Дюран (Duran,

1983) отметил, что изменение тестов не дает перспективного решения этих проблем среди испаноязычных студентов; скорее здесь нужны прямые исследования и решения. Тем не менее тестовые показатели следует интерпретировать с учетом всей информации о биографических переменных, действующих как модераторы в индивидуальных случаях. Более того, любые обобщения в отношении испаноязычных американцев должны принимать в расчет возможные различия между подгруппами: пуэрториканцами, мексиканцами и т. д.

Систематическая ошибка интерцепта. Даже когда тест дает одинаковые коэффициенты валидности для двух групп, он может тем не менее обнаружить систематическую ошибку интерцепта. Интерцепт — это отрезок, отсекаемый линией регрессии на координатной оси. Тест показывает систематическую ошибку интерцепта, если систематически занижает или завышает предсказуемое выполнение критерия для конкретной группы. Вернемся к случаю 1 на рис. 6–5, в котором выборки меньшинства и большинства показывают идентичные регрессии. В этих условиях нет ни ошибки наклона, ни ошибки интерцепта. Когда группы значимо различаются по средним показателям теста, они обнаруживают соответствующие различия и в выполнении критериальной деятельности. В случае 3 линии регрессии двух групп имеют один и тот же наклон, но разные интерцепты. Здесь у группы меньшинства *(A)* более высокий интерцепт, чем у группы большинства *(B)*, т. е. линия регрессии меньшинства пересекает ось *Y* выше, чем линия регрессии большинства. Несмотря на то что коэффициенты валидности, вычисленные в каждой группе, равны, любой тестовый показатель *(X)* будет соответствовать в этих двух группах различным критериальным показателям, что показано на рисунке точками Y_A и Y_B. Таким образом, один и тот же тестовый показатель имеет разное прогнозирующее значение для этих групп.

Психологи, которых беспокоит возможная несправедливость тестов по отношению к представителям разных меньшинств, как раз и имеют в виду ситуацию, представленную случаем 3. Заметим, что в этом случае большинство превосходит группы меньшинств по результатам тестирования, но и большинство, и меньшинства одинаково хорошо выполняют критериальную деятельность. Тем самым отбор всех претендентов на основе критического тестового показателя, установленного для группы большинства, несправедливо дискриминировал бы меньшинство. При этих условиях применение регрессии, построенной по данным большинства, к обеим группам приводит к *недооценке* предсказываемого выполнения критерия представителями группы меньшинства. Подобная ситуация, по-видимому, может возникнуть, когда значительная часть дисперсии показателей теста не имеет отношения к прогнозируемому критерию и характеризует функции, в которых большинство превосходит данное меньшинство. Полный анализ выполняемой работы и удовлетворительная валидность тестов служат мерами, предохраняющими от выбора такого теста.

Проблема систематической ошибки интерцепта имеет самое непосредственное отношение к тому, что в народе называют «честностью теста» (*test fairness*). Хотя выражения «честность теста» и «необъективность теста» (в смысле систематической ошибки) употребляются как равнозначные и настолько широкие, что охватывают все аспекты тестирования культурных меньшинств, уже стало привычным отождествлять честность (или нечестность) теста с систематической ошибкой интерцепта. Такого употребления придерживались авторы «Единых нормативов по методам отбора наемных работников» (*Uniform Guidelines on Employee Selection Procedures*, 1978). В разделе «Честность» (14 B) основное положение сформулировано следующим образом:

> В тех случаях, когда для представителей одной расовой, половой или этнической группы типично получать в ходе отбора более низкие показатели по сравнению с представителями другой группы, и эти различия в показателях не отражаются на различиях в мере выполнения работы, использование данной процедуры отбора может несправедливо лишать возможностей членов группы, получающей относительно низкие показатели.

Однако эмпирические исследования существующей практики использования тестов либо свидетельствовали об отсутствии значимой систематической ошибки интерцепта, либо чаще выявляли слабую тенденцию *противоположного* направления, представленную случаем 4 на рис. 6–5. Здесь у группы большинства *(В)* более высокий интерцепт, чем у группы меньшинства *(А)*. При этих условиях применение регрессии и критического показателя, построенным по данным большинства, к обеим группам ведет к *переоценке* предсказываемого выполнения критериальной деятельности членами группы меньшинства и тем самым к несправедливой дискриминации группы большинства. Такие результаты были получены в исследованиях предсказания успеваемости в колледже (Breland, 1979; Duran, 1983; Zeidner, 1987) и юридической школе (Linn, 1975), успешности освоения программ подготовки специалистов в сухопутных и военно-воздушных силах (M. A. Gordon, 1953; Maier, & Fuchs, 1973; C. W. Shore, & Marion, 1972), а также широкого множества производственных критериев (см. обзор в Hunter et al., 1984).

Как было доказано математически, случай 4 (рис. 6–5) имеет место, если две группы различаются по одной или нескольким *дополнительным переменным (additional variables)*, которые положительно коррелирует как с тестом, так и с критерием (Linn, & Werts, 1971; Reilly, 1973). Несколько завышенный прогноз является статистическим артефактом учета только одного предиктора зараз. С добавлением предикторов к тестовой батарее это завышение уменьшается, — факт, который получил эмпирическое подтверждение в различных совокупностях, от студентов-юристов и конторских служащих до питомцев детских садов (см. Hunter et al., 1984).

Интересно отметить, что те же результаты были получены при сравнении групп, различавшихся по образовательному или социоэкономическому уровню. Армейская классификационная батарея завышала прогнозируемое выполнение программы обучения военной специальности для тех, кто был отчислен из старших классов школы, и занижала его для выпускников колледжей (Maier, 1972). Аналогично этому, заниженный прогноз успеваемости по результатам тестов академических способностей имел место для студентов, у которых профессиональное положение отцов было достаточно высоко, и завышенный прогноз — для студентов, чьи отцы занимали более низкое профессиональное положение (Hewer, 1965). Во всех этих исследованиях сравнение групп с высокими и низкими тестовыми показателями либо вообще не обнаруживало значимого различия в интерцепте, либо выявляло небольшую систематическую ошибку в пользу группы с более низкими показателями по тестам.

Модели принятия решений для честного использования тестов. Постепенно фокус исследований начал перемещаться от оценивания систематической ошибки тестов к разработке стратегий отбора для честного использования тестов в работе с культурными меньшинствами. Если стратегия отбора строится исходя из регрессионной модели (см. модель Клири), иллюстрация которой дана на рис. 6–5, людей будут выбирать (при приеме в колледж, на работу и т. д.) исключительно на основе их прогнозируе-

мых показателей критериальной деятельности. Такая стратегия будет максимизировать общий результат критериальной деятельности, безотносительно к другим целям процесса отбора. Согласно этой стратегии, честным использованием тестов при отборе будет их использование, опирающееся только на наилучшую оценку выполнения критерия для каждого конкретного человека.

Предлагали и другие модели принятия решения, имевшие своей целью отбор большей доли лиц из группы с низкими тестовыми показателями. Эта цель соответствует задаче, которую обычно определяют в таких терминах, как «позитивные действия»[1] или ослабление «неблагоприятного воздействия» процесса отбора. Во время внедрения этих альтернативных моделей казалось, что они руководствуются методами, совершенно отличными от тех, которые предполагает регрессионная модель.[2] Однако позднее было показано, что все эти модели можно выразить в виде вариантов одной общей модели (Darlington, 1971; Gross, & Su, 1975; Petersen, & Novick,1976). Различия между ними допускают объяснение исходя из ценностных суждений, имплицитно содержащихся в каждой модели. Роль ценностей в стратегиях принятия решений уже обсуждалась в этой главе (см. рис. 6–2). Напомним, что приписывание относительной *полезности* результату каждого решения требует оценки степени благоприятности или неблагоприятности такого результата. Эти субъективные оценки, вместе с вероятностью каждого результата, используют при вычислении общей ожидаемой полезности (*EU*) стратегии.

Основанный на теории принятия решений анализ честного использования тестов показал, что предложенные модели различаются своим определением честности, — в той мере, в какой они имплицитно придают различную ценность принятию и отверганию потенциальных успехов и неудач внутри совокупностей меньшинств и большинства. Модели ожидаемой полезности выражают основные социальные ценности в явном виде. Этот подход обязывает открыто формулировать оценки полезностей, которые невозможно получить статистическими методами, ибо они предполагают широкое обсуждение и последовательное приближение к балансу конфликтующих целей (N. S. Cole, & Moss, 1989; Darlington, 1976; Messick, 1989). К числу таких целей относятся обеспечение равенства возможностей для всех людей, максимизация успеха и продуктивности, увеличение демографического разнообразия рабочей силы (по крайней мере, для некоторых профессий) и расширение преференциального режима для групп, поставленных в невыгодное положение несправедливыми действиями в прошлом.

Наконец, следует особо подчеркнуть, что статистические корректировки тестовых баллов, критических показателей и формул предсказания вряд ли можно рассматривать как перспективные средства исправления последствий социальной несправедливости. Использование статистических манипуляций, маскирующих различия пока-

---

[1] В Америке политическая программа, направленная на ликвидацию расовой дискриминации. — *Примеч. науч. ред.*

[2] Литература по разнообразным моделям принятия решений для честного использования тестов весьма обширна и в большинстве своем посвящена техническим вопросам. Что касается краткого изложения характерных особенностей и последствий применения разных моделей, см. Bond (1981), Dunnette & Borman (1979, pp. 497–500), Gross & Su (1975, p. 350–351), C. R. Reynolds (1982). Более полные пояснения можно найти в Hunter & Schmidt (1976) и Hunter et al. (1977).

зателей путем установления отдельных норм для подгрупп или рас[1], по всей видимости, все же наносит вред конкретным людям вследствие распределения их по рабочим местам или образовательным программам, для которых они не подходят из-за отсутствия необходимых навыков или знаний. Результатом часто становится плохая работа или учеба, что не только сказывается на Я-концепции человека и его отношении к делу, но может способствовать поддержанию социального стереотипа в отношении представителей некоторой культурной или этнической группы как плохих работников, нерадивых студентов и т. п. Более конструктивные решения предлагаются в рамках других подходов, уже обсуждавшихся в этой главе. Один из них показан на примере тестирования комплекса способностей и стратегий распределения, позволяющих максимально использовать многообразные паттерны способностей, сформировавшиеся под влиянием разных культурных истоков. Более широкое рассмотрение релевантных черт личности, мотивации и аттитюдов также облегчает прогнозирование трудовых или учебных достижений. Еще один подход основан на применении адаптивных программ типа индивидуализированного обучения. Чтобы такие программы максимально соответствовали индивидуальным особенностям, тесты должны как можно полнее и точнее определять наличный уровень развития необходимых способностей у каждого их участника. Общие, комплексные модели принятия решений создают условия для объединения разных подходов и систем ценностей и для оценивания результирующей эффективности каждого решения.

[1] См., например, D. C. Brown (1994), L. S. Gottfredson (1994), Sackett & Wilk (1994).

7 АНАЛИЗ ЗАДАНИЙ

Знакомство с основными понятиями и методами анализа заданий, равно как и с другими аспектами конструирования тестов, может помочь пользователям в оценке выпускаемых тестов. Кроме того, анализ заданий особенно важен при составлении неформальных, локальных тестов, наподобие вариантов опросов или контрольных работ, которые учитель готовит для использования в своем классе. Знание ряда общих принципов и правил составления эффективных заданий, вместе с овладением наиболее простыми статистическими методами их анализа, может существенно повысить качество таких классных тестов и сделать их пригодными для применения даже в небольших группах.

В заданиях может анализироваться как их качественная сторона, т. е. их содержание и форма, так и количественная, т. е. их статистические свойства. Качественный анализ включает рассмотрение содержательной валидности (обсуждавшейся в главе 5) и оценивание заданий с точки зрения эффективных методов их составления. Количественный анализ предполагает главным образом измерение трудности и различительной способности заданий. Валидность и надежность любого теста в конечном счете зависят от характеристик входящих в него заданий. Высокую валидность и надежность можно заложить в тест заранее, на этапе анализа заданий. Тест можно значительно улучшить, удаляя, добавляя, заменяя или пересматривая отдельные задания.

Анализ заданий позволяет сократить тест и в то же время повысить его валидность и надежность. При прочих равных условиях более длинный тест валиднее и надежнее короткого. Влияние увеличения или сокращения теста на коэффициент надежности обсуждалось в главе 4, где также была приведена формула Спирмена—Брауна для оценивания этого влияния. Эти предполагаемые (оцениваемые с помощью формулы Спирмена—Брауна) изменения надежности теста происходят в тех случаях, когда изымаемые задания равноценны оставшимся или когда добавляемые задания равноценны уже имеющимся в его составе. Аналогичные изменения валидности теста возникают в результате удаления или добавления заданий равноценной валидности. Все такие оценки изменения надежности или валидности относятся к увеличению или сокращению теста путем *случайного* отбора заданий, проводимого без их анализа. Когда же сокращение теста идет за счет исключения наименее удовлетворительных заданий, короткий тест может оказаться более валидным и надежным, чем его первоначальная полная версия.

Трудность заданий

Процент справившихся с заданием. Для большинства целей тестирования трудность задания определяется в единицах процента (или доли) лиц, давших на него правильный ответ. Чем легче задание, тем выше этот процент. Слово, значение которого правильно указало 70 % выборки стандартизации ($p = 0{,}70$), считается более легким, чем слово, которое знают только 15 % ($p = 0{,}15$). Обычно задания располагаются в порядке нарастания трудности, так, чтобы тестируемый начинал с относительно легких заданий и затем переходил ко все более сложным. Такое расположение дает тестируемому больше уверенности в своих силах и снижает вероятность того, что он, затратив много времени на задания, которые для него слишком трудны, пропустит те, которые ему по силам.

В процессе конструирования теста основным оправданием измерения трудности заданий служит требование подбора заданий подходящего уровня сложности. Большинство стандартизованных тестов способностей создается с расчетом на получение для каждого тестируемого как можно более точной оценки его уровня достижений в области конкретной способности. Согласно такой цели, если ни один тестируемый не справляется с предложенным заданием, то оно оказывается просто лишним грузом в данном тесте. То же можно сказать и о заданиях, с которыми справляются все. Ни те ни другие не дают никакой информации об индивидуальных различиях. А поскольку такие задания не влияют на изменчивость тестовых показателей, они не вносят никакого вклада в надежность или валидность теста. Чем ближе трудность задания к 1,00 или к 0, тем менее дифференцированную информацию о тестируемых можно получить с его помощью. И наоборот, чем ближе уровень трудности задания к 0,50, тем больше разграничений можно сделать с его помощью. Предположим, что из 100 тестируемых 50 справились и 50 не справились с заданием ($p = 0{,}50$). Это задание позволяет нам провести попарное различие между каждым, кто справился и кто не справился с ним, что дает $50 \times 50 = 2500$ парных сравнений, или двоичных единиц (битов) различительной информации. Задание, с которым справляется 70 % тестируемых, дает $70 \times 30 = 2100$ битов информации; когда с заданием справляется 90 % тестируемых, оно дает $90 \times 10 = 900$ битов информации; когда же с ним справляются все 100 %, оно дает $100 \times 0 = 0$ битов информации, т. е. абсолютно неинформативно. Те же соотношения остаются в силе и для более трудных заданий, с которыми справляется менее 50 % тестируемых.

Тогда, в целях максимизации различительной способности теста, казалось бы, следует подбирать все его задания на уровне трудности 0,50. Решение, однако, осложняется тем обстоятельством, что в рамках одного теста задания имеют тенденцию коррелировать друг с другом. Чем однороднее тест, тем выше эти корреляции. В предельном случае, если бы все задания имели уровень трудности 0,50 и полностью коррелировали между собой, с каждым заданием в итоге справились бы одни и те же 50 человек из 100. Следовательно, половина тестируемых получила бы высший показатель, а другая половина — нулевой. По причине корреляции заданий между собой, их лучше всего отбирать таким образом, чтобы уровень трудности отдельных заданий имел некоторый умеренный разброс, но в среднем составлял 0,50. Кроме того, чем выше взаимокорреляции заданий (или корреляции заданий с суммарным показателем), тем шире должен быть их разброс по уровню трудности.

Еще одно соображение, принимаемое в расчет при выборе подходящего уровня трудности заданий, касается вероятности угадывания ответа в заданиях с множественным выбором. Чтобы учесть возможность выбора определенной частью тестируемых правильного ответа путем угадывания, требуемая доля правильных ответов устанавливается выше той, которую можно было бы ожидать в случае задания со свободным ответом. Например, для задания с выбором из 5 вариантов средняя доля правильных ответов должна составлять примерно 0,69 (Lord, 1952).

Интервальные шкалы. Процент лиц, справившихся с заданием, выражает его трудность в единицах порядковой шкалы, т. е. правильно указывает ранговый порядок, или относительную трудность заданий. Если, к примеру, с заданиями 1, 2 и 3 справляется соответственно 30 %, 20 % и 10 % тестируемых, то мы можем заключить, что задание 1 — самое легкое, а задание 3 — самое трудное из этих трех. Но мы не можем утверждать, что различие в трудности между заданиями 1 и 2 то же, что и между заданиями 2 и 3. Равные разности процентов соответствовали бы равным различиям в трудности заданий только при прямоугольном распределении, в котором случаи равномерно распределены по всему диапазону. Эта проблема аналогична той, с которой мы встретились в связи с процентильными показателями, также основанными на процентах случаев. Напомним из главы 3, что процентильные показатели не представляют собой равных единиц и меняются по величине при переходе от центра к краям распределения (рис. 3–4).

Если исходить из нормального распределения свойства, измеряемого любым данным заданием, то уровень трудности задания можно выразить в единицах шкалы равных интервалов, пользуясь таблицей значений плотности нормального распределения. В главе 3 мы видели, например, что при нормальном распределении примерно 34 % случаев попадает в интервал между средним и величиной, равной $+1\sigma$ или -1σ (рис. 3–3). С учетом этой информации рассмотрим рис. 7–1, показывающий уровень трудности задания, с которым справились 84 % тестируемых. Поскольку правой («верхней») части распределения соответствуют лица, справившиеся с заданием, а левой («нижней») — не справившиеся с ним, эти 84 % включают в себя всю правую половину (50 %) и часть (34 %) левой половины (50 + 34 = 84). Следовательно, это задание

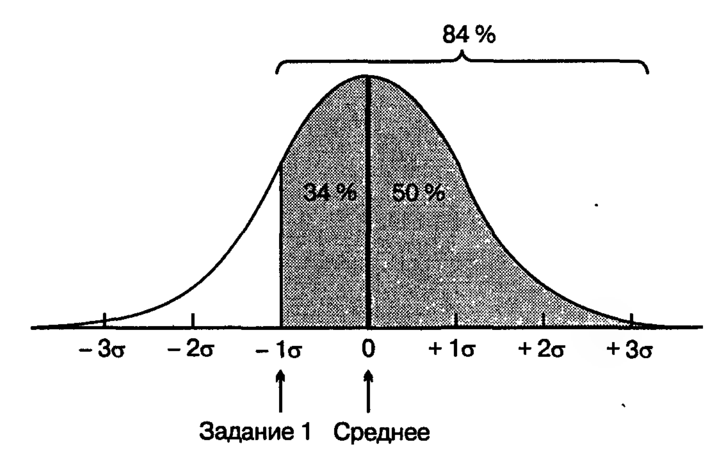


Рис. 7–1. Соотношение между процентом справившихся с заданием и его сложностью, выраженной в единицах нормального распределения

(по уровню трудности) находится на 1σ *ниже* среднего, как и показано на рис. 7–1. Задание, выполненное только 16 % тестируемых, находилось бы на 1σ *выше* среднего по своей сложности, так как в область справа от этой точки попадает 16 % случаев (50 – 34 = 16). Задание, с которым справились точно 50 % тестируемых, находилось бы в точке, соответствующей среднему нормального распределения, и получило бы нулевое значение по этой шкале. Таким образом, задания выше среднего уровня сложности оцениваются положительными величинами, а задания ниже среднего уровня сложности — отрицательными величинами. Стандартную оценку трудности, соответствующую любому проценту справившихся с заданием лиц, можно найти по таблице значений плотности нормального распределения, имеющейся в любом типовом учебнике по статистике.

Абсолютное шкалирование по Тёрстоуну. Индексы трудности задания, выраженные в процентах или единицах нормальной кривой (т. е. в единицах стандартного отклонения), ограничены диапазоном проявления изучаемой способности в выборке, на которой они вычислялись. Для некоторых целей тестирования, однако, нужна мера трудности заданий, пригодная для разных выборок, варьирующих по уровню способности. Например, в образовательных тестах достижений бесспорным преимуществом была бы возможность сравнивать в единой шкале показатель ребенка при переходе из класса в класс на протяжении какого-то периода обучения. При всем этом явно нереальной задачей было бы пытаться шкалировать входящие в них задания, предназначенные для всех классов, путем предъявления этих заданий какой-то одной группе, поскольку одни из них оказались бы слишком трудными, а другие — слишком легкими почти для каждого члена такой группы.

Другим примером могут служить крупномасштабные программы тестирования, требующие множества эквивалентных форм для разновременного проведения теста, такие как программы приема в высшие учебные заведения. Эта проблема рассматривалась в главе 3 постольку, поскольку она затрагивает интерпретацию совокупных показателей, получаемых с помощью таких инструментов, как Тест академической оценки (*Scholastic Assessment Test*). Предложенное решение проблемы состояло в том, чтобы использовать фиксированную эталонную группу для определения нулевой точки и единиц шкалы, а затем все последующие показатели переводить в такую шкалу. Это преобразование требует набора анкерных, или связующих заданий, которые включаются в состав тестов, проводимых в любой паре групп. Такие задания составляют минитест в том смысле, что они являются репрезентативным — по форме и содержанию — отображением полного теста. Для разных пар групп могут использоваться свои, отличные от других, наборы связующих заданий. Каждая новая форма теста связывается с одной или двумя более ранними его формами, а те, в свою очередь, с другими формами посредством цепи таких минитестов, тянущейся назад вплоть до исходной эталонной группы.

Тем же общим методом можно воспользоваться для измерения трудности отдельных заданий в единой шкале, применимой к любому числу взаимосвязанных групп. Соответствующая статистическая процедура, называемая абсолютным шкалированием, была разработана Тёрстоуном (Thurstone, 1925, 1947а) и широко использовалась при разработке тестов (например, Donlon, 1984). По существу, эта процедура состоит из двух шагов. Сначала мы находим шкальные оценки заданий отдельно в каждой группе, преобразуя процент справившихся с каждым из них людей в единицы *y*-отклоне-

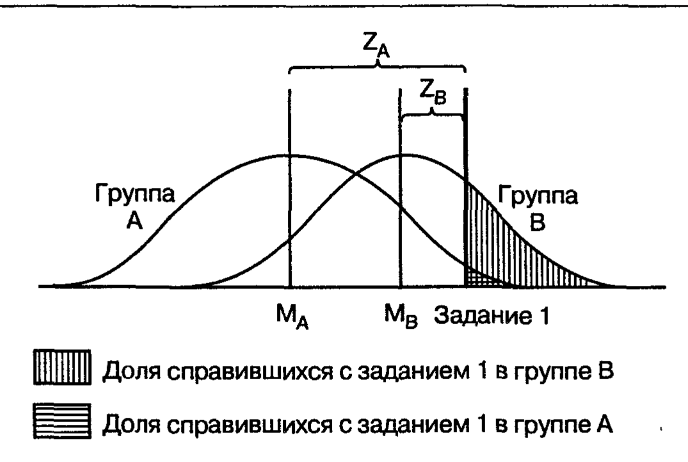
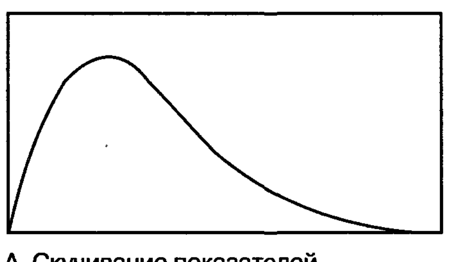


Рис. 7–2. *Z*-оценки, показывающие относительную сложность одного задания в группах А и В

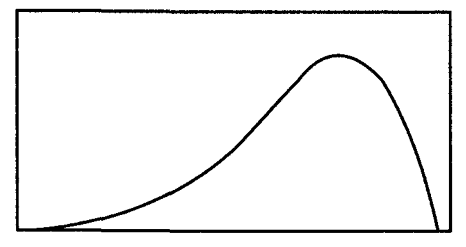
ния (т. е. стандартного отклонения) нормальной кривой или *z*-оценки. Затем мы переводим все эти шкальные оценки в соответствующие оценки для *одной* из обследованных групп, принятой нами за стандартную, или эталонную группу. В качестве эталонной может быть выбрана любая группа, скажем, протестированная первой, самая младшая, средняя по уровню выполнения заданий или какая-то другая подходящая для целей тестирования группа. Все, что требуется, — это набор общих, анкерных заданий, которые предъявляются двум или большему числу групп и шкалируются внутри каждой группы.

Шкальные оценки одних и тех же заданий в двух (или более) группах используют для определения отношения между группами и позволяют преобразовывать все оценки трудности заданий при переходе от одной группы к другой. Это отношение схематически проиллюстрировано на рис. 7–2, показывающем *у*-отклонение (т. е. величину *z*) одного и того же задания (*i*) в двух соседних группах, *А* и *В*. С этим заданием (*i*) в группе *В* справляется большая доля лиц, чем в группе *А*. Поэтому его *у*-отклонение от своего группового среднего меньше в группе *В* (z_B), чем в группе *А* (z_A). Соответствующие величины z_A и z_B для всех *общих заданий* обеспечивают базу для формулы перевода, посредством которой *все задания*, предъявленные в группе *В*, можно переоценить по уровню трудности применительно к группе *А*, и наоборот. Простую номограмму для приближенного перевода оценок легко получить, построив график зависимости z_A от z_B (проведя через соответствующие точки прямую линию). Эту линию можно затем использовать для нахождения значений z_A для всех других заданий, предъявляемых группе *В*.

Ту же процедуру перевода оценок можно распространить на любое число групп, работая с парами соседних, частично перекрывающихся групп. Например, в тесте, рассчитанном на учащихся 1–8-х классов, оценочную шкалу для восьмиклассников можно преобразовать в шкалу для семиклассников, а шкалу для семиклассников — в шкалу для шестиклассников, и так далее, до первого класса. Группы из соседних классов обычно обладают достаточной степенью сходства (или перекрытия), чтобы обеспечить использование значительной части теста в целях согласования оценок. Однако любой отдельный школьный класс будет иметь разные общие части теста с ближайшим старшим и младшим классами.



А. Скучивание показателей на нижнем конце шкалы



В. Скучивание показателей на верхнем конце шкалы

Рис. 7–3. Асимметрия кривых распределения

Распределение тестовых показателей. Трудность теста в целом, разумеется, напрямую зависит от трудности заданий, из которых он состоит. Полную проверку трудности всего теста применительно к популяции, для которой он создавался, обеспечивает распределение его суммарных показателей. Если выборка стандартизации представляет собой репрезентативный срез такой популяции, то можно ожидать приблизительно нормального распределения его показателей.

Предположим, однако, что эмпирическая кривая распределения явно отличается от теоретической нормальной кривой своей асимметрией, или скошенностью, как это показано на рис. 7–3 (*А* и *В*). Первое из этих распределений (*А*), с выраженной правосторонней асимметрией (т. е. с преобладанием в выборке низких тестовых показателей), свидетельствует о слишком высоком уровне теста для данной группы, в котором не достает легких заданий, чтобы должным образом дифференцировать тестируемых в левой (нижней) области распределения. В силу этого лица, показатели которых при нормальных условиях тестирования имели бы значительный разброс, получают в этом тесте показатели близкие или равные нулю, — отсюда и пик в нижней части шкалы. Эта искусственная «штабелевка» показателей схематически проиллюстрирована на рис. 7–4, где нормально распределенная по уровню способности группа дает скошенное распределение показателей по конкретному тесту. Распределение с противоположной, левосторонней асимметрией показано на рис. 7–3 (*В*). Здесь показатели группируются преимущественно на верхнем конце шкалы, что свидетельствует о чересчур низком потолке трудности в данном тесте. Такого рода скошенное распределение наблюдается, например, когда тест, предназначенный для общей популяции, дается выборке студентов или аспирантов, часть которых получает почти абсолютные, предельные показатели. С помощью такого теста невозможно измерить индивидуальные различия среди наиболее способных студентов или аспирантов в группе. Если бы в тест были включены более трудные задания, некоторые из них наверняка набрали бы больше баллов, чем позволяет получить данная версия теста.

Когда распределение тестовых показателей, полученное на выборке стандартизации, заметно отличается от нормального, уровень трудности теста обычно корректируют до тех пор, пока кривая распределения не оказывается примерно нормальной. В зависимости от типа отклонений от нормального распределения добавляются более легкие или более трудные задания, первоначальные задания изымаются или видоизменяются, меняется их положение в шкале или пересматриваются приписываемые определенным ответам веса, используемые при вычислении показателя по данному тесту. Все эти корректировки продолжают до тех пор, пока не получают распределение

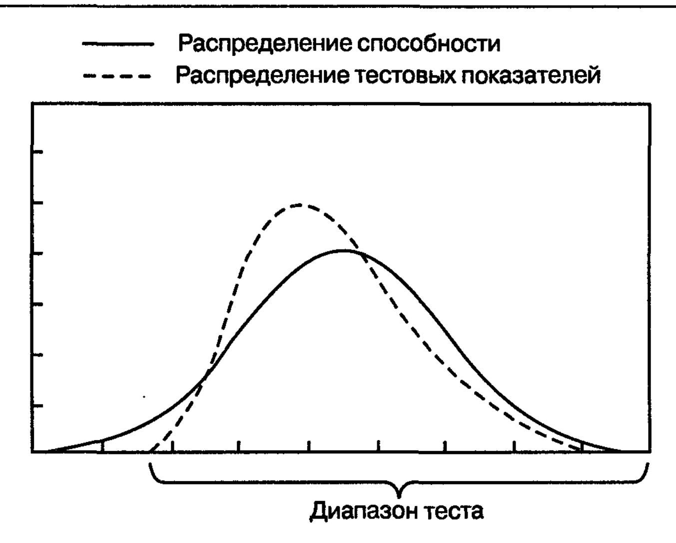


Рис. 7–4. Асимметрия распределения тестовых показателей, возникающая в результате недостаточного количества легких заданий в составе теста

показателей, имеющее хотя бы грубое сходство с нормальным. При этих условиях наиболее вероятный показатель, получаемый большинством тестируемых, соответствует примерно 50 % правильно выполненных заданий. Тому, кто не знаком с методами конструирования психологических тестов, 50 %-ный результат может показаться поразительно низким. Иногда именно на этом основании проводящему тестирование специалисту высказывают возражения против установленного им якобы слишком низкого норматива прохождения данного теста. Или же делается вывод, будто протестированная группа оказалась исключительно слабой. Несостоятельность подобных мнений сразу становится очевидной, если принять во внимание все те процедуры, которые используются при разработке психологических тестов. Такие тесты сознательно конструируются и модифицируются с таким расчетом, чтобы они давали средний показатель, примерно соответствующий 50 % правильно выполненных заданий. Только таким путем можно добиться максимальной дифференциации обследуемых лиц на всех уровнях способности, получаемой с помощью данного теста. При среднем, составляющем приблизительно 50 % правильно выполненных заданий, создается максимальная возможность получить нормальное распределение с широким рассеянием индивидуальных показателей на обоих его краях.[1]

Увязывание трудности заданий с целью тестирования. Стандартизованные психологические тесты обычно создавали с целью добиться максимально возможной на всех уровнях дифференциации тестируемых. Наше обсуждение трудности заданий

[1] В действительности нормальная кривая обеспечивает более тонкое различение на краях, чем в середине шкалы. Для получения равной различительной способности шкалы во всех ее точках потребовалось бы прямоугольное распределение. Однако нормальная кривая предпочтительнее с точки зрения последующего статистического анализа показателей, поскольку многие современные статистические методы основываются на распределении, близком к нормальному. По этой и другим причинам составители большинства тестов, предназначенных для широкого использования, вероятно, будут еще какое-то время ориентироваться на нормальную кривую.

до сих пор относилось к тестам именно такого рода. Однако при конструировании тестов специального назначения выбор уровня трудности заданий, так же как и оптимальной формы распределения тестовых показателей, зависит от типа искомой дифференциации. Так, в тестах, предназначенных для целей отсеивания, следует применять задания, уровень трудности которых приближается к заданному коэффициенту отбора. Например, чтобы отобрать 20 % группы тестируемых с самыми высокими показателями, лучше всего использовать задания, группирующиеся около $p = 0{,}20$ (или несколько выше, чтобы учесть возможность угадывания). Так как в тесте отсеивания никакой дифференциации *внутри* принятых или непринятых групп не требуется, время тестирования используется наиболее эффективно в том случае, когда задания группируются около значения критического показателя. Отсюда, например, вытекает, что если тест предназначен для отбора из совокупности студентов кандидатов на получение стипендии, то его задания должны быть значительно труднее среднего уровня заданий для такой популяции. Аналогично, если отбираются плохо успевающие ученики для коррекционной программы обучения, задания желательно выбирать намного легче среднего уровня.

Другой пример выбора уровня трудности заданий исходя из специальных целей тестирования можно найти в области тестирования владения знаниями, умениями и навыками, или, короче, владения предметом или деятельностью. Напомним (см. главу 3), что такое тестирование часто сочетается с предметно-ориентированным тестированием. Если назначение теста — установить, овладел ли индивидуум как следует основными, существенными элементами того или иного умения или усвоил ли он знания, необходимые для перехода к следующему этапу обучения, то трудность заданий должна быть на уровне 0,8–0,9. При этих условиях мы могли бы ожидать, что большинство экзаменуемых справится почти со всеми заданиями. Таким образом, самые легкие задания (даже те, с которыми справляются 100 % тестируемых), изымаемые из обычного стандартизованного теста из-за их низкой различающей способности, и есть те задания, которые включаются в тест владения предметом или деятельностью. Подобным же образом тест, проводимый перед началом очередного этапа обучения, с тем чтобы определить, не приобрел ли уже кто-то из учеников те умения и навыки, которым их собираются учить, будет давать очень низкий процент правильных ответов по каждому заданию. В этом случае задания с очень низким и даже нулевым p не следует выбрасывать из теста, поскольку они выявляют то, чему еще предстоит научиться.

Из приведенных примеров хорошо видно, что уровень трудности заданий зависит от назначения теста. Хотя в большинстве ситуаций тестирования максимум информации об уровне деятельности каждого индивидуума дают задания средней трудности, группирующиеся около $p = 0{,}50$, решение о трудности заданий нельзя принимать шаблонно, без учета того, как предполагается использовать тестовые показатели.

Различительная способность заданий

Выбор критерия. Под различительной способностью задания понимают ту степень, с какой оно правильно дифференцирует тестируемых по поведению, для измерения которого и предназначен данный тест. В тех случаях, когда тест в целом можно оценить посредством критериальной валидизации, входящие в него задания также

могут оцениваться и отбираться на основе их связей с тем же внешним критерием. Этим путем особенно часто шли при разработке некоторых тестов личности и интересов, обсуждаемых в главах 13 и 14. Кроме того, этот метод обычно используют при выборе вопросов для включения в биографические вопросники, которые в типичном случае охватывают разнородное собрание сведений о происхождении и жизненном пути конкретных лиц. Применительно к измерительным инструментам этого типа мы не располагаем никаким априорным основанием для классификации ответов на правильные и неправильные или для приписывания им весовых коэффициентов, кроме сравнения с критериальным статусом лиц, дающих эти ответы. Из первоначального банка заданий (вопросов) сохраняются те, которые лучше всего дифференцируют обследуемых лиц, отнесенных к различным критериальным категориям, таким как различные профессии или психиатрические синдромы. Часто критериальные группы состоят из достигших успеха и потерпевших неудачу в университетском курсе, программе профподготовки или конкретном виде работы.

В предметно-ориентированном тестировании уровня знаний, умений и навыков, обсуждавшемся в главе 3, задания могут оценивать путем сравнения выполнения каждого из них лицами, различающимися объемом полученного обучения в соответствующей области (Panell, & Laabs, 1979; L. A. Shepard, 1984). Обычной практикой является сравнение долей лиц, давших правильные ответы на задания до и после прохождения курса обучения. Поскольку эти тесты используют для определения того, достигли ли обучаемые заданного уровня владения предметом или деятельностью, индивидуальные различия в результатах при однократном проведении теста сведены к минимуму. При этих условиях внутренний анализ заданий (предполагающий их сравнение друг с другом) не будет иметь смысла и поэтому нужен внешний критерий, такой как объем обучения в конкретной области.

В других типах тестов достижений, как и во многих тестах способностей, различительная способность заданий обычно исследуется по отношению к суммарному показателю самого теста.[1] Для образовательных тестов достижений внешний критерий в типичных случаях недоступен. Что касается тестов способностей, растущее внимание исследователей к конструктной валидности и методам ее установления делает суммарный показатель по тесту вполне уместным критерием для отбора заданий. На начальных этапах разработки теста суммарный показатель обеспечивает первое приближение к мере изучаемой способности, черты или конструкта.

Рассмотрим более подробно следствия выбора заданий на основе внешнего критерия и на основе суммарного тестового показателя. Первый путь ведет к максимизации валидности теста относительно внешнего критерия, а второй — к максимизации внутренней согласованности или однородности теста. При определенных условиях эти два подхода могут приводить к противоположным результатам: задания, выбираемые по соображениям внешней валидности, оказываются как раз теми заданиями, которые отбрасываются исходя из соображений внутренней согласованности. Предположим, что предварительная форма теста академических способностей состоит из 100 арифметических и 50 словарных заданий. Чтобы произвести отбор заданий из этой исход-

[1] Корреляции «задание — тест» всегда несколько завышены из-за совместного действия дисперсии ошибок и специфической дисперсии конкретного задания и теста, частью которого оно является. Для корректировки этого эффекта «часть — целое» имеются специальные формулы (Guilford & Fruchter, 1978, p. 165–167).

ной совокупности с целью повышения внутренней согласованности теста, необходимо будет вычислить некий показатель согласования между выполнением каждого задания и суммарным показателем по 150 заданиям. Очевидно, что такой показатель, в общем, будет выше для арифметических, чем для словарных заданий, так как суммарный показатель основан на в два раза большем числе арифметических заданий. Если мы захотим сохранить 75 «лучших» заданий в окончательной форме этого теста, то большинство из них, по всей вероятности, окажутся арифметическими. Но с точки зрения внешнего критерия академической успеваемости, словарные задания, возможно, были бы более валидными предикторами, чем арифметические. Если дело обстоит именно так, то анализ заданий привел бы к снижению, а не повышению валидности теста.

Практика отбрасывания заданий, имеющих низкие корреляции с суммарным показателем, дает нам способ повышения однородности или «очищения» теста. Благодаря применению этой процедуры сохраняются задания с наибольшими средними интеркорреляциями. Данный метод отбора заданий будет повышать валидность теста только в тех случаях, когда первоначальная совокупность заданий измеряет одно-единственное свойство и когда это свойство присутствует в критерии или оцениваемом конструкте. Однако некоторые типы тестов измеряют комбинацию свойств, требуемых сложным критерием. В таком случае очищение теста может привести к сужению зоны охвата тестом его критерия и тем самым к снижению валидности.

Отбор заданий с целью максимизации критериальной валидности теста можно уподобить отбору тестов для получения наибольшей валидности батареи. Напомним (глава 6), что вклад теста в валидность батареи тем больше, чем выше его корреляция с критерием и чем ниже корреляция с другими тестами батареи. Если этот принцип применить к отбору заданий, то наиболее удовлетворительными будут задания с самыми высокими показателями внешней валидности и самыми низкими коэффициентами внутренней согласованности. Так, задание, имеющее высокую корреляцию с внешним критерием, но относительно низкую — с суммарным показателем теста, было бы предпочтительнее задания, имеющего высокую корреляцию и с критерием, и с тестом в целом, ибо первое задание, по-видимому, измеряет некоторый аспект критерия, не охватываемый в должной мере оставшейся частью теста.

Казалось бы, при отборе заданий можно использовать те же методы, что и при выборе тестов для включения в батарею. В частности, можно было бы вычислить корреляцию каждого задания с критерием и со всеми остальными заданиями. Лучшим заданиям, отобранным таким путем, можно было бы затем приписать веса на основе построенного уравнения регрессии. Такая процедура, однако, неосуществима и теоретически несостоятельна. Дело не только в большом объеме необходимых для этого вычислений, но и в том, что корреляции между заданиями сильно зависят от колебаний выборки и найденные по ним коэффициенты регрессии были бы слишком неустойчивы, чтобы на них можно было основывать отбор заданий. Есть и более серьезное возражение против такой процедуры: получившийся в результате тест оказался бы настолько неоднородным по содержанию, что это исключило бы всякую возможность смысловой интерпретации тестового показателя.

Валидность относительно внешнего критерия и внутренняя согласованность являются важными целями конструирования теста. Относительное значение, придаваемое каждой из них, меняется в зависимости от характера и назначения теста. Применительно ко многим задачам тестирования удовлетворительным компромиссным ре-

шением будет сгруппировать относительно однородные задания в отдельные тесты или субтесты, каждый из которых охватывает какой-то один аспект внешнего критерия. Тем самым широта охвата достигается за счет разнообразия тестов, каждый из которых дает более или менее однозначный показатель, а не за счет разнородности заданий в рамках одного теста. При таком подходе задания с низкими индексами внутренней согласованности не отбрасывались бы, а выделялись в особые группы. В результате этого внутри каждого субтеста или группы заданий можно было бы достичь довольно высокой внутренней согласованности.

Статистические индексы различительной способиости задания. Поскольку обычно регистрируется лишь факт выполнения или невыполнения задания, измерение различительной способности задания, как правило, связано с соотнесением дихотомической переменной (задания) и непрерывной переменной (критерия). В некоторых ситуациях критерий тоже может быть дихотомической переменной, как в случае окончания или отчисления из колледжа, успеха или неудачи в работе. Кроме того, непрерывный критерий в целях анализа всегда можно преобразовать в дихотомический.

Было разработано свыше 50 индексов различительной способности задания, которые и в настоящее время используют при конструировании тестов. Одно из различий между ними относится к применимости этих индексов к дихотомическим или непрерывным мерам. Кроме того, среди индексов, применимых к дихотомическим переменным, одни предполагают непрерывность и нормальное распределение измеряемого с помощью теста свойства, которое подвергается искусственной дихотомизации при обработке результатов тестирования, тогда как другие основаны на предположении об истинной дихотомии изучаемого свойства. Другое различие касается связи трудности задания с различительной способностью. Некоторые индексы оценивают различительную способность задания независимо от его трудности, а некоторые дают более высокую оценку различительной способности заданий, уровень трудности которых приближается к 0,50, и более низкие оценки для крайне легких и крайне трудных заданий.

Независимо от способа получения и исходных допущений большинство индексов различительной способности задания дают весьма сходные результаты (Oosterhof, 1976). Хотя числовые значения индексов могут различаться, на их основе сохраняются или отвергаются в основном одни и те же задания. В действительности, колебание данных о различительной способности задания от выборки к выборке в целом больше, чем при использовании различных методов получения таких данных.

Использование контрастных групп. Распространенный метод анализа заданий — сравнение долей выполнивших задание в двух контрастных по выполению критерия группах. Когда критерий измеряется в непрерывной шкале (как в случае годовых оценок, оценок работы руководителями, показателей производительности труда или суммарных показателей по определенному тесту), верхняя *(В)* и нижняя *(Н)* критериальные группы формируются из лиц, занимающих положение на соответствующих краях распределения. Очевидно, что чем ближе к краям распределения будут эти группы, тем резче будет выражено различие. Однако использование предельно контрастирующих групп, представленных, скажем, верхними и нижними 10 % распределения, снизило бы надежность результатов из-за малого числа используемых случаев. При нормальном распределении оптимальная точка, в которой эти два условия уравновеши-

ваются, достигается при верхних и нижних 27 % распределения (T. L. Kelley, 1939). Когда распределение более плоско, чем нормальная кривая, оптимальный процент несколько больше 27 % и равен почти 33 % (Cureton, 1957b). В случае малых групп — таких, как обычный класс, — ошибка выборки настолько велика, что можно рассчитывать только на грубые статистические оценки. Поэтому здесь не приходится заботиться о точном проценте случаев в двух контрастных группах. Приемлема любая цифра между 25 и 33 %.

При разработке стандартизованных тестов используются большие и нормально распределенные выборки, и в этом случае обычно работают с верхними и нижними 27 % распределения критериальных показателей. Многие таблицы и номограммы, облегчающие вычисление индексов различительной способности заданий, составлены на основе допущения о соблюдении «правила 27 %». По-видимому, распространение быстродействующих компьютеров позволит со временем заменить различные вспомогательные приемы, разработанные для облегчения анализа заданий, более точными и совершенными методами. Современная вычислительная техника позволяет анализировать результаты всей выборки, не ограничиваясь верхним и нижним краями распределения.

Упрощенный анализ заданий в случае малых групп. Поскольку анализ заданий часто проводится при работе с малыми группами, например с учащимися одного класса, отвечающими на контрольный вопросник, рассмотрим прежде всего простую процедуру, особенно подходящую для такой ситуации. Предположим, в классе всего 60 человек, из которых отобрано 20 учеников (33 %) с самыми высокими и 20 (33 %) — с самыми низкими тестовыми показателями. Разложим листки с ответами на три стопки, принадлежащие верхней (В), средней (С) и нижней (Н) группе. Теперь нам нужно определить, сколько правильных ответов в каждой из этих групп было дано на каждый вопрос. Для этого выпишем в столбик номера вопросов, оставив справа место для трех колонок, которые обозначим буквами В, С и Н. Возьмем из стопки В любой листок и в колонке В проставим палочки против тех вопросов, на которые данный ученик ответил правильно. Это нужно проделать для каждого из 20 листков группы В, затем для 20 листков группы С и, наконец, для всех листков группы Н. Подсчитаем теперь палочки и запишем результаты для каждой группы так, как это показано в табл. 7–1 (для краткости в ней приведены цифры только по первым семи вопросам). Приблизительный индекс различительной силы любого из вопросов находится вычитанием числа учеников, правильно ответивших на него в группе Н, из числа учеников, правильно ответивших на него в группе В. Эти разности (В—Н) приведены в последней колонке табл. 7–1. На основе тех же исходных данных можно получить меру трудности вопроса, для чего нужно сложить число справившихся с каждым вопросом во всех трех критериальных группах (В + С + Н).

Анализ табл. 7–1 выявляет 4 сомнительных задания, которые заслуживают последующего рассмотрения или обсуждения в классе. Два вопроса, 2-й и 7-й, были выделены потому, что один из них слишком легок (56 из 60 учеников ответили на него правильно), а другой слишком труден (всего 5 правильных ответов). Вопросы 4-й и 5-й, хотя и удовлетворительны с точки зрения уровня трудности, тем не менее обнаруживают отрицательную и нулевую различительную способность соответственно. К этой категории мы также отнесли бы любые вопросы с очень малыми положительными значениями разности (В — Н), примерно от 3 и менее единиц для сравниваемых

Таблица 7–1

Упрощенная процедура анализа заданий: число лиц, давших правильный ответ в каждой критериальной группе

Задание (вопрос)	В (20)	С (20)	Н (20)	Трудность (В + С + Н)	Различительная способность (В – Н)
1	15	9	7	31	8
2	20	20	16	56*	4
3	19	18	9	46	10
4	10	11	16	37	–6*
5	11	13	11	35	0*
6	16	14	9	39	7
7	5	0	0	5*	5
•					
•					
•					
•					
75					

* Задания, выбранные для последующего обсуждения

групп примерно того же размера. Имея дело с большими группами, можно ожидать и больших различий (В—Н), возникающих случайно при выполнении задания, не обладающего различительной способностью.

Цель анализа заданий теста, подготовленного учителем, состоит в выявлении дефектов как в самом тесте, так и в преподавании. Одного обсуждения сомнительных заданий с классом часто достаточно для того, чтобы обнаружить проблему. Если вопрос сформулирован неудачно, его можно перестроить или вовсе изъять при последующем тестировании. Обсуждение, однако, может обнаружить, что вопрос составлен правильно, но у учеников нет надлежащего понимания данной темы. В этом случае тема может быть разобрана заново и пояснена подробнее. При отыскании менее очевидного источника затруднений часто полезно провести дополнительный анализ (см. табл. 7–2) хотя бы тех вопросов, что были отобраны для обсуждения. В табл. 7–2 приводится число учеников из групп В и Н, выбравших каждый из пяти вариантов ответа на эти вопросы.

Хотя вопрос 2 и был включен в табл. 7–2, мы мало что можем узнать о нем из приведенных здесь данных о частоте ошибочных ответов, поскольку неправильный выбор сделали лишь 4 ученика из группы Н и никто — из группы В. Однако обсуждение этого вопроса с учениками, возможно, поможет определить, действительно ли вопрос слишком легок и не представляет особой ценности, или какой-то недостаток формулировки позволяет сразу же находить правильный ответ, или же, наконец, это полезный вопрос, но относится к хорошо проработанной с учителем и прочно усвоенной теме занятий. В первом случае вопрос, видимо, следует изъять, во втором — переформулировать, а в третьем — оставить без изменения.

Данные по вопросу 4 показывают, что третий вариант ответа содержит в себе нечто такое, что заставляет 9 учеников из группы В предпочесть его правильному (второму) варианту. В чем здесь дело, нетрудно установить, попросив этих учеников обосновать свой выбор. Ошибки в ответах на вопрос 5, видимо, объясняется неудачностью формулировки либо самого вопроса, либо варианта правильного ответа, так как ошибоч-

Таблица 7–2

Анализ ответов на отдельные вопросы

Задание (вопрос)	Группа	Варианты ответов				
		1	2	3	4	5
2	В	0	0	0	**20**	0
	Н	2	0	1	**16**	1
4	В	0	**10**	9	0	1
	Н	2	**16**	2	0	0
5	В	2	3	3	**11**	2
	Н	1	3	3	**11**	2
7	В	**5**	3	5	4	3
	Н	0	5	8	3	4
•						
•						
•						

Примечание. Правильные варианты ответов выделены жирным шрифтом

ные выборы учащихся равномерно распределились по четырем вариантам ложного ответа. Вопрос 7 необычно труден: 15 человек из группы В и вся группа Н ответили на него неправильно. Несколько больший выбор третьего (ложного) варианта в данном случае указывает на его внешнюю привлекательность, особенно для легче вводимых в заблуждение членов группы Н. Аналогично отсутствие правильных ответов (вариант 1) в группе Н говорит о том, что плохо осведомленному ученику эта альтернатива на первый взгляд кажется ошибочной. Разумеется, оба эти свойства желательны для хорошего тестового задания. Обсуждение в классе могло бы показать, что вопрос 7 — это хороший вопрос, относящийся, однако, к теме, усвоенной лишь несколькими учениками данного класса.

Индекс различительной способности. Если число справившихся с определенным заданием в верхней (В) и нижней (Н) критериальных группах выразить в процентах, разность между ними дает нам индекс различительной способности задания, интерпретируемый независимо от размера выборки, на которой он был получен. Этот индекс неоднократно обсуждался в психометрической литературе (см., например, Ebel, 1979; A. P. Johnson, 1951; Oosterhof, 1976) и обозначался то как $U - L$[1], то как *ULI* или *ULD*, а то и просто *D*. Несмотря на свою простоту, этот индекс, как было показано, хорошо согласуется с другими, более сложными мерами различительной способности задания (Engelhart, 1965; Oosterhof, 1976). Вычисление *D* можно проиллюстрировать на примере данных, приведенных в табл. 7–1. Сначала число лиц, справившихся с каждым заданием в группах В и Н, переводится в проценты. Разность между соответствующими процентами и есть индекс различительной способности (*D*), значения которого для семи анализируемых нами заданий приведены в табл. 7–3. *D* может принимать любое значение между ±100. Если все члены группы В справились и никто из группы Н не справился с заданием, то $D = 100$. И наоборот, если группа Н справилась, а группа В не справилась с заданием, то $D = -100$. Если же процент справившихся с заданием в обеих группах одинаков, то $D = 0$.

[1] Первые буквы английских слов *Upper* (верхний) и *Lower* (нижний). — *Примеч. науч. ред.*

Таблица 7–3

Вычисление индекса различительной способности задания

Задание (вопрос)	Процент справившихся с заданием		Индекс различительной способности (*D*)
	Группа В	Группа Н	
1	75	35	40
2	100	80	20
3	95	45	50
4	50	80	–30
5	55	55	0
6	80	45	35
7	25	0	25

Примечание. Использованы данные из табл. 7–1

Как и другие индексы различительной способности заданий, индекс *D* зависит от трудности задания, но в отличие от них обнаруживает смещение в пользу промежуточных уровней трудности. В табл. 7–4 приведены максимально возможные значения *D* для заданий с различным процентом правильных ответов. В тех случаях, когда 100 % или 0 % всей выборки справились с заданием, никакого различия в процентах справившихся с этим заданием в группах В и Н просто не может быть, — и потому *D* = 0. С другой стороны, если с заданием справились 50 % членов выборки, не исключено, что все они принадлежат к группе В, и тогда *D* = 100 (100 – 0 = 100). Если же справившихся оказалось 70 %, то максимальное значение, которое индекс *D* мог бы принять в этом случае, можно пояснить следующим образом: (В) 50/50 = 100 %; (Н) 20/50 = = 40 %; *D* = 100 – 40 = 60. Напомним, что для большинства целей тестирования предпочтение отдается заданиям, уровень трудности которых близок к 0,50. Поэтому индексы различительной способности, принимающие максимальные значения при этом уровне трудности, часто более других подходят для отбора заданий.

Таблица 7–4

Связь максимальной величины индекса *D* с трудностью задания

Процент справившихся с заданием	Максимальная величина *D*
100	0
90	20
70	60
50	100
30	60
10	20
0	0

Коэффициент ϕ. Многие индексы различительной способности заданий выражают связь между заданием и критерием в виде коэффициента корреляции. Одним из них является коэффициент ϕ (фи). Вычисляемый по четырехпольной таблице, ϕ основан на соотношении долей справившихся и не справившихся с заданием в верхней (В) и нижней (Н) критериальных группах. Подобно всем коэффициентам корреляции, ϕ принимает значения в интервале от +1,0 до –1,0 и предполагает подлинную дихотомию как ответов на задание, так и критериальной переменной. Следовательно,

он применим лишь к тем дихотомическим условиям, при которых был найден, и не может быть обобщен на какие-то глубинные, скрывающиеся за ними отношения между измеряемыми данным заданием свойствами и критерием. Как и индекс *D*, коэффициент ϕ принимает наибольшие значения для заданий средних уровней трудности, когда дихотомия близка к соотношению 50 : 50.

Уровень значимости коэффициента ϕ нетрудно определить благодаря его связи и с критерием χ^2, и с *Z*-критерием (критическим отношением). С помощью последнего можно найти минимальное значение ϕ, достигающее статистической значимости на уровне 0,05 или 0,01, по следующим формулам:

$$\phi_{.05} = \frac{1{,}96}{\sqrt{N}},$$

$$\phi_{.01} = \frac{2{,}58}{\sqrt{N}}$$

В этих формулах *N* — суммарное число испытуемых в обеих критериальных группах. Так, если бы группы *В* и *Н* содержали по 50 человек, то *N* было бы равно 100, и минимальное значение коэффициента ϕ, значимое на уровне 0,05, равнялось бы $1{,}96 : \sqrt{100} = 0{,}196$. Следовательно, любое задание с коэффициентом ϕ, равным или превышающим 0,196, коррелировало бы с критерием на уровне значимости 0,05.

Бисериальная корреляция. В качестве последнего примера широко используемой меры различительной способности задания можно рассмотреть бисериальную корреляцию (r_{bis}), отличающуюся от ϕ в двух важных отношениях. Во-первых, r_{bis} предполагает непрерывное и нормальное распределение свойств, лежащих в основе дихотомической формы ответа на задание и критериальной переменной. Во-вторых, r_{bis} как мера связи между заданием и критерием не зависит от трудности задания.

Для оценки бисериальной корреляции нужно знать средние критериальных показателей справившихся и не справившихся с заданием и соответствующее *SD*, вычисленное по показателям всех членов критериальной группы, а также долю лиц, справившихся (либо не справившихся) с заданием в этой группе. Формулы для вычисления r_{bis} приведены в большинстве учебников по статистике (например, Guilford, & Fruchter, 1978, pp. 304–306). Стандартную ошибку r_{bis} можно вычислить с помощью простой формулы, включающей выражения из формулы для вычисления r_{bis}. Следует добавить, что наличие вычислительной техники позволяет сразу получать значения r_{bis} и их стандартных ошибок.

Теория «задание — ответ»

Регрессия «задание — тест». Трудность и различительную способность задания можно одновременно отобразить в виде линии регрессии «задание — тест». В целях иллюстрации рассмотрим гипотетический тест из 12 заданий, требующих коротких ответов в свободной форме, наподобие словарных тестов в проводимых индивидуально шкалах интеллекта. В табл. 7–5 приведены доли лиц с разным суммарным баллом по этому тесту, ответивших правильно на каждое из двух заданий. Эти же данные представлены в виде графиков на рис. 7–5.

Уровень трудности каждого задания можно определить как его 50 %-ный порог, так же как это обычно делается при установлении сенсорных порогов в психофизике. Это сделано на рис. 7–5 с помощью простейших геометрических построений: из точек пересечения кривых двух заданий с горизонтальной линией, соответствующей 50 % правильных ответов, опускают два перпендикуляра на ось абсцисс, по которой отложены суммарные тестовые показатели (баллы). Из этих построений хорошо видно, что у тех, кто набрал по этому тесту в сумме примерно 8 баллов, шансы справиться с заданием 7 равны 50 : 50, а у набравших примерно 10 баллов такие же шансы справиться с заданием 11. На различительную силу каждого задания указывает крутизна соответствующей кривой: чем круче кривая, тем выше корреляция выполнения задания с суммарным показателем по тесту и больше величина индекса различительной способности задания. Судя по внешнему виду кривых, различительная способность заданий 7 и 11 примерно одинакова.

Изучение регрессий «задание—тест», подобных изображенным на рис. 7–5, дает возможность наглядно представить, насколько эффективно работает то или иное задание теста. Такие графики не только объединяют информацию о трудности и различительной способности задания, но также дают полную картину отношений между выполнением каждого задания и суммарным тестовым показателем. Например, задание 7 обнаруживает инверсию, поскольку те, кто набрал в сумме 10 баллов, справляются с этим заданием лучше тех, кто набрал 11 баллов по данному тесту. Когда подобные результаты получены на малой выборке, этой инверсией можно было бы пренебречь; однако она иллюстрирует вид информации, которую могут выявить данные такого анализа заданий.

Несмотря на очевидные достоинства, такие графики являются довольно грубыми и мало пригодны для математической обработки, точной оценки и строгого отбора заданий. Этот подход послужил отправной точкой для разработки весьма тонких и сложных типов анализа заданий, которые начали завоевывать внимание в 1970-х и начале 1980-х гг. Причину их растущей популярности, безусловно, следует искать в

Таблица 7–5

Гипотетические данные для построения регрессии «задание—тест»

Суммарный показатель (балл)	Доля правильных ответов	
	Задание 7	Задание 11
12	1,00	0,95
11	0,82	0,62
10	0,87	0,53
9	0,70	0,16
8	0,49	0,05
7	0,23	0,00
6	0,10	0,00
5	0,06	0,00
4	0,03	0,00
3	0,00	0,00
2	0,00	0,00
1	0,00	0,00

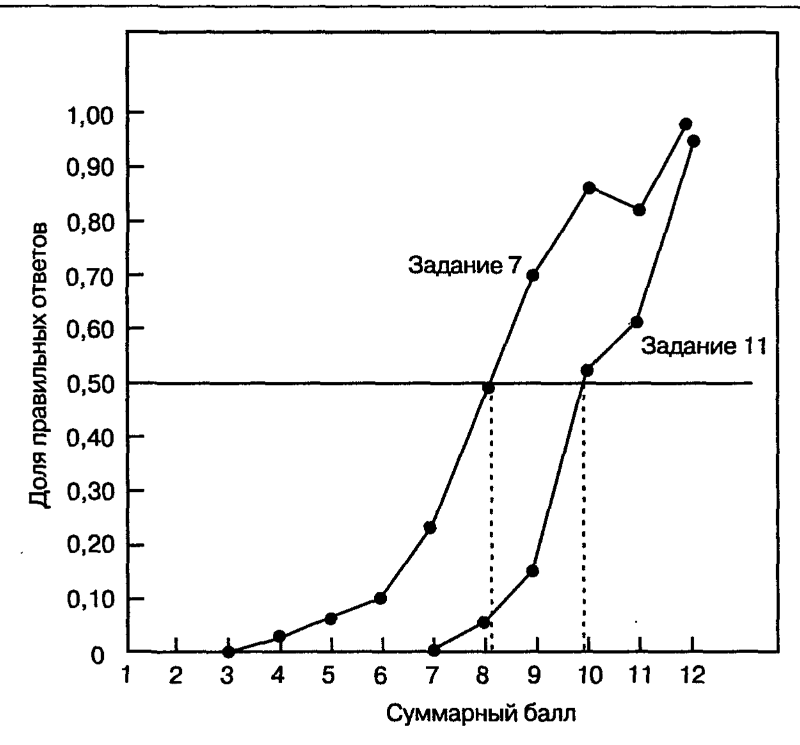


Рис. 7-5. Регрессия «задание—тест» для заданий 7 и 11 (по данным табл. 7-5)

стремительном расширении доступа к быстродействующим компьютерам, без которых связанные с такими типами анализа вычислительные задачи потребовали бы несоразмерных затрат времени и средств. С составлением компьютерных программ для целого ряда предложенных моделей анализа заданий, практическое применение этих тонких методов стало легко осуществимым. Важнейшие особенности этого подхода будут охарактеризованы в следующих разделах.

Теория «задание — ответ» (*IRT*): основные черты.[1] Рассматриваемый математический подход — теория «задание — ответ» — также известен под названиями «теория латентных черт» и «теория характеристических кривых задания» (*item characteristic curve theory* или, сокращенно, *ICC* теория). Главная особенность этого подхода состоит в том, что выполнение задания ставится в связь с оценкой величины «латентной черты» респондента, обозначаемой греческой буквой (тэта). В этом контексте под «латентной чертой» понимается статистический конструкт, за которым не стоит никакой психологической или физиологической сущности, обладающей независимым существованием. В когнитивных тестах латентной чертой обычно называют измеряемую тестом способность (*ability*). Суммарный показатель по тесту часто принимают за начальную оценку такой способности.

[1] Ясный обзор методологии *IRT* и ее приложений см. в Hambleton et al. (1991). Обзоры технических аспектов *IRT* и ее критические оценки можно найти в Hambleton (1989), Drasgow & Hulin (1990). О внедрении *IRT* в психометрику см. Lord (1980), D. J. Weiss (1983), D. J. Weiss & Davidson (1981).

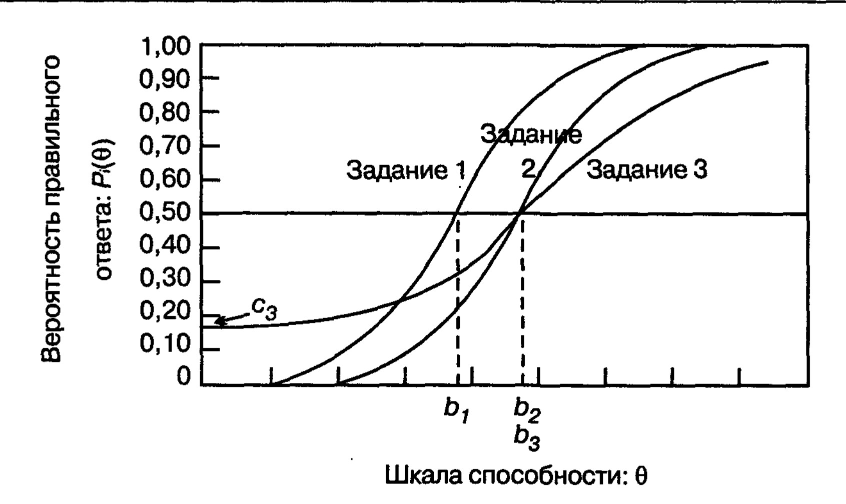


Рис. 7–6. Гипотетические характеристические кривые для трех заданий

Характеристические кривые заданий строятся на основе математически выведенных функций, а не по эмпирическим данным, используемым при построении регрессионных кривых «задание—тест». В различных моделях *IRT* используются разные математические функции, так как эти модели основаны на разных наборах допущений. В одних моделях — это интегральные кривые нормального распределения; в других — логистические функции, позволяющие использовать некоторые математически удобные свойства логарифмических отношений. Вообще, применение различных моделей этого рода дает по существу сходные результаты, при условии, что лежащие в их основе допущения не нарушаются в конкретных ситуациях. На рис. 7–6 изображены характеристические кривые для трех гипотетических заданий. Осью абсцисс задана шкала способности (θ), оцениваемой по суммарному тестовому показателю и другой информации об ответах на тест в конкретной выборке. Ось ординат дает значения P_i (θ) — вероятности правильного ответа на i-е задание как функции от положения респондента на шкале способности (θ). Эта вероятность находится по данным о доле респондентов, отнесенных к разным уровням изучаемой способности, которые справились с i-м заданием.

В полной, трехпараметрической модели каждая *ICC* описывается тремя параметрами, выведенными математически из эмпирических данных. Параметр различающей мощности (или различительной способности) задания (a_i) свидетельствует о наклоне кривой. Он обратно пропорционально связан с тем расстоянием, на которое нужно переместиться по континууму способности (θ), чтобы повысить P_i (θ). Чем больше величина a_i, тем круче наклон кривой. На рис. 7–6 задания 1 и 2 имеют одинаковую величину a_i, или различающую мощность; задание 3 характеризуется меньшим a_i, так как его кривая поднимается медленнее. Параметр трудности задания (b_i) соответствует точке на оси способности, в которой вероятность правильного ответа, P_i (θ), равна 0,50. Из рисунка хорошо видно, что задания 2 и 3 имеют одинаковый параметр b_i и, значит, одинаковую трудность, а задание 3 легче и, следовательно, требует меньшей способности для достижения вероятности правильного ответа P_i (θ) = = 0,50. Модели *IRT* для заданий с множественным выбором часто включают третий

параметр — так называемый параметр угадывания (c_i).[1] Он отображает вероятность случайного появления правильного ответа. При использовании заданий с множественным выбором даже у обследуемых с самими низкими уровнями способности вероятность дать правильный ответ выше нуля. На рис. 7–6 это видно на примере задания 3, чья асимптота снизу проходит значительно выше нуля.

В типичных случаях для вычисления оценок параметров задания и оценок способности используют итеративные методы или, как их еще называют, методы последовательного приближения; аппроксимации повторяются до тех пор, пока оценки не становятся устойчивыми. В добавление к получению математически уточненных индексов трудности и различительной способности заданий методы *IRT* дают ряд других преимуществ. Важной особенностью этого подхода является исследование надежности и ошибки измерения при помощи *информационных функций заданий (item information functions)*. Эти функции, вычисляемые для каждого задания, служат надежной опорой при выборе заданий в процессе конструирования теста. Информационная функция задания учитывает все его параметры и показывает его эффективность как средства измерения на различных уровнях способности.

Наиболее широко разрекламированный вклад моделей *IRT* имеет отношение к получаемым с их помощью результатам, которые не зависят от характера выборки, что в специальной литературе описывается как *инвариантность параметров задания (invariance of item parameters)*. Основная идея теории «задание — ответ» как раз и состоит в том, что параметры задания не должны изменяться при их вычислении в группах, различающихся по уровню способности. Кроме того, это означает, что как группы, так и отдельных людей можно тестировать с помощью разных наборов заданий, которые соответствуют их уровням способности, а их показатели можно сравнивать напрямую. Тестовый показатель каждого конкретного человека основывается не только на количестве, но и на заранее установленном уровне трудности выполненных им правильно заданий.

Когда предполагается тестирование множества различных выборок, единственный возможный способ — работать с большой совокупностью или банком заданий, предварительно откалиброванных на большой случайной выборке. В тех случаях, когда диапазон способности очень широк, как это имеет место в серии тестов достижений, охватывающих все ступени школьного обучения, для преодоления разрывов между группами необходимо использовать общие задания (называемые по-разному: анкерными, согласующими или калибровочными). После того как задания в полной совокупности будут откалиброваны, любое их подмножество можно применять для тестирования любой группы или отдельного человека, а полученные показатели — сравнивать между собой.

Другие модели *IRT*. В предыдущем разделе мы рассматривали трехпараметрическую модель. Двухпараметрические модели, с опущенным параметром случайного ответа (c_i), применяют в тех случаях, когда влиянием угадывания правильных ответов на выполнение теста можно пренебречь. Однопараметрическая модель, основанная только на учете трудности(b_i) набора заданий, была разработана Рашем (Rasch,

---

[1] Некоторые исследователи рекомендуют называть c_i просто асимптотой снизу (*lower-asymptote*) или случайным параметром *ICC*, потому что трехпараметрические модели трактуют c_i как величину, не зависящую от способности, тогда как в действительности угадывание является функцией способности.

1966; см. также Andersen, 1983) и, в последующем, развита и поддержана рядом исследователей (например, Wright, 1977; Wright, & Stone, 1979). Эта модель основана на предположении о том, что как угадывание, так и изменение различительной силы задания на разных уровнях способности не оказывают существенного влияния на выполнение теста. На практике, при конструировании теста, сторонники модели Раша часто отбрасывают именно те задания, которые нарушают это предположение. Кроме того, нередко заявлялось, что модели *IRT* являются «робастными» в статистическом смысле, а значит допускающими, в определенных границах, нарушение разных предположений без искажения результатов. Разумеется, выяснить это можно только путем эмпирической проверки.

Рассматриваемые до сих пор модели предполагают *одномерность (unidimensionality)* теста или, иначе говоря, исходят из допущения, что ответы на задание можно объяснить одним свойством или одной чертой. В общем, предположение одномерности может в достаточной мере удовлетворяться, если выполнение теста зависит от единственной преобладающей черты, даже когда другие черты менее значительным образом, но все же сказываются на результатах тестирования. Были также сконструированы более общие модели, применимые к многомерным тестам, однако они требуют и более трудоемких вычислительных процедур. Кроме того, были разработаны различные модификации моделей для обработки ответов с несколькими градациями (а не только дихотомических) (Samejima, 1969) или для анализа различных вариантов ответов в заданиях со множественным выбором (Bock, 1972).

Современное состояние *IRT*. В отношении достоинств альтернативных моделей *IRT* все еще продолжаются широкие дискуссии. Математически получаемые на основе этих моделей оценки требуют гораздо более серьезной проверки, причем не только с помощью моделированных данных и машинного моделирования, но и на реальных данных. Инвариантность параметров задания особенно нуждается в широком исследовании в реальных ситуациях. Например, одни и те же задания могут потребовать различной смеси способностей при выполнении их лицами с различным жизненным и профессиональным опытом или же одним человеком на разных стадиях научения. Если посмотреть с другой стороны, то для анализа данных с помощью моделей *IRT* уже сейчас доступно большое количество разнообразных компьютерных программ (см., например, Hambleton, 1989, р. 171–172); однако эти программы постоянно меняются в результате переоценки, пересмотра и замены.[1]

Несмотря на продолжающийся рост теоретического и методологического разнообразия в этой области, использование методов *IRT* в практической разработке тестов неуклонно возрастает. Технические приемы *IRT* быстро включаются в состав как вновь создаваемых тестов, так и пересмотренных версий широко используемых тестовых батарей, разработанных коммерческими издательствами. В качестве примеров можно назвать Калифорнийские тесты достижений (*California Achievement Tests*) и Комплексные тесты основных навыков (*Comprehensive Tests of Basic Skills*), а также Дифференциальные шкалы способностей, характеристика которых дана в главе 8. *IRT*

---

[1] Самый известный и свежий пример — программа *ASCAL* для двух- и трехпараметрической логистической *IRT* калибровки, распространяемая корпорацией *ASC* (адрес указан в приложении Б). Уместно указать и на недавнюю разработку обобщенной линейной теории «задание—ответ» (*GLIRT*), из которой можно выводить различные модели *IRT* и которая допускает приспособление к разным форматам заданий (Mellenbergh, 1994).

особенно подходит для некоторых недавно появившихся типов тестирования, таких как компьютеризованное адаптивное тестирование (*КАТ*), рассматриваемое в главе 10. В ходе такого тестирования каждый тестируемый может отвечать на разные наборы заданий, однако все ответы оцениваются по единой шкале (Wainer et al., 1990). Важным приложением *IRT* является применение этого подхода в долгосрочном проекте разработки *КАТ* версии Батареи профессиональной пригодности Вооруженных сил США (*Armed Services Vocational Aptitude Battery*) (Wiskoff, & Schratz, 1989).

Анализ заданий тестов скорости

Независимо от того, важна ли скорость для измеряемой функции, индексы заданий, вычисленные по скоростному тесту, могут вводить в заблуждение. Если не считать заданий, при выполнении которых никто или почти никто из обследуемых не испытывал недостатка времени, эти индексы будут отражать не столько действительную трудность или различительную силу того или иного задания, сколько его *положение (position)* в данном тесте. С заданиями, появляющимися в тесте позднее, справится сравнительно меньшая доля общей выборки, поскольку лишь немногие успеют до них добраться. Каким бы легким ни было задание, если оно расположено в конце теста скорости, оно будет выглядеть трудным. Если, скажем, вопрос об имени тестируемого поместить в конце скоростного теста, то процент лиц, ответивших на него, был бы весьма низким.

Подобным же образом завышаются индексы различительной способности тех заданий, к выполнению которых не все тестируемые успевают приступить. Поскольку более опытные испытуемые обычно работают быстрее, у них больше шансов добраться до заданий, находящихся в конце теста скорости. Таким образом, независимо от характера самого задания некоторая корреляция между ним и критерием будет обнаруживаться просто потому, что оно появляется ближе к концу теста скорости.

Чтобы избежать некоторых из этих затруднений, можно было бы ограничить анализ каждого задания только данными тех лиц, которые достигли соответствующего задания в тесте. Это решение, однако, нельзя считать вполне удовлетворительным, если число лиц, сумевших добраться до анализируемого задания, мало. Такая процедура сопряжена с использованием быстро сокращающегося числа тестируемых, вследствие чего результаты по последним заданиям могут оказаться ненадежными. Кроме того, лица, выполнившие такие задания, вероятно, будут представлять собой селективную выборку, не сопоставимую с более широкой выборкой, использованной для анализа ранних заданий. Как уже отмечалось, испытуемые, работающие быстро, часто и более опытны. Таким образом, более поздние задания будут анализироваться на выборке лиц из верхней части распределения. Одним из эффектов такого селективного фактора могло бы оказаться снижение видимого уровня трудности более поздних заданий, поскольку процент справившихся с заданием в селективной выборке был бы выше, чем в полной выборке. Отметим, что в данном случае ошибка обратна той, которая появляется при вычислении процента справившихся с заданием по данным всей выборки. В последнем случае происходит искусственное завышение видимой трудности заданий.

Влияние вышеупомянутой процедуры на индексы различительной способности заданий не столь очевидно, но тем не менее реально. Замечено, например, что некото-

рые из тестируемых с низкими показателями склонны спешить при выполнении теста, отвечая на задания почти случайным образом в своем стремлении опробовать их все в рамках отведенного времени. Среди получивших высокие показатели эта тенденция выражена гораздо меньше. В результате выборка, на которой производится анализ поздно появляющегося в тесте задания, нередко включает нескольких весьма слабых респондентов, выполняющих это задание на уровне случайности, и большее число опытных и быстрых респондентов, чьи ответы обычно оказываются правильными. В такой группе корреляция задания и критерия, вероятно, будет выше, чем в более репрезентативной выборке. С другой стороны, без таких случайных респондентов выборка, на которой анализируются расположенные в конце теста задания, охватывает относительно узкий диапазон способности. При этих условиях индексы различительной способности более поздних заданий, вероятно, будут ниже, чем в том случае, когда они вычисляются на всей выборке.

Ожидаемое влияние скорости на индексы трудности и различительной способности заданий проверялось опытным путем как для случаев, когда статистики задания вычислялись по данным полной выборки (Wesman, 1949), так и для случаев, когда они вычислялись по данным только тех лиц, которые пытались выполнить данное задание (Mollenkopf, 1950a). Во втором из этих двух исследований сопоставимым группам старшеклассников давали две формы вербального теста и две формы математического теста. Каждая из двух форм состояла из одних и тех же заданий, но их начальные и конечные серии в этих формах менялись местами. Каждая форма предъявлялась в жестких (условия скорости) и свободных (условия возможностей) временных рамках. Такой план эксперимента позволял проводить разнообразные сравнения между формами тестов и временными условиями. Результаты ясно показали, что положение задания в тестах скорости влияло на его индексы трудности и различительной способности. Когда одно и то же задание предъявлялось позднее в скоростном тесте, оно выполнялось бóльшим процентом испытуемых, пытавшихся его решить, и давало более высокую корреляцию с критерием.

Трудности, возникающие в ходе анализа заданий скоростных тестов, в принципе аналогичны тем, о которых говорилось в главе 4 в связи с надежностью тестов скорости. Были предложены различные — как эмпирические, так и статистические, — способы преодоления этих трудностей. Одним из эмпирических решений было увеличение лимита времени для группы, на которой проводится анализ заданий. Такое решение приемлемо, если только сама скорость не является важным аспектом измеряемой тестом способности. Однако помимо технических проблем, связанных с конкретными тестами, необходимо иметь в виду, что данные, получаемые в ходе анализа заданий скоростных тестов, сомнительны сами по себе и требуют тщательной проверки.

Перекрестная валидизация

Смысл перекрестной валидизации. Важно, чтобы валидность теста определялась на выборке испытуемых, отличной от той, на которой производился отбор заданий. Это независимое определение валидности всего теста называется перекрестной, или кросс-валидизацией. На любом коэффициенте валидности, найденном по выборке, применявшейся для отбора заданий, будут сказываться ошибки случайного отбора испытуемых, приводя к искусственному завышению его величины. Фактически, при

таких обстоятельствах высокий коэффициент валидности можно было бы получить даже в том случае, когда тест совершенно не обладает валидностью в предсказании конкретного критерия.

Предположим, что в выборке из 100 студентов-медиков были выделены 30 человек с самыми высокими и 30 с самыми низкими баллами по медицинским дисциплинам, которые составили контрастные критериальные группы. Если теперь эти две группы сопоставить по ряду свойств, фактически не имеющих отношения к успеваемости в медицинском колледже, то, несомненно, будут обнаружены те или иные случайные различия. Так, в верхней критериальной группе может оказаться больше выпускников частных школ и рыжеволосых студентов. Если бы нам пришло в голову приписывать каждому человеку по дополнительному баллу за окончание частной школы и за рыжий цвет волос, то средний показатель оказался бы, несомненно, выше в верхней, чем в нижней критериальной группе. Однако это не является доказательством валидности выбранных нами прогнозирующих признаков, так как такой процесс валидизации содержит круг в доказательстве. Оба прогнозирующих признака выбраны в первую очередь на основе случайной вариации, которая характеризует данную выборку. И *те же* случайные различия ответственны за появление среднегрупповых различий в суммарных показателях. Однако при проведении теста в другой выборке случайные различия в количестве окончивших частные школы и рыжих, скорее всего, исчезнут или изменят знак, и следовательно, валидность показателей резко снизиться.

Эмпирический пример. Классическое доказательство необходимости перекрестной валидизации дает раннее исследование, проведенное с тестом чернильных пятен Роршаха (Kurtz, 1948). Чтобы выяснить, мог ли этот тест чем-то помочь при отборе кандидатов на должность коммерческого директора агентства по страхованию жизни, он был проведен на 80 таких директорах. Они были тщательно отобраны из нескольких сотен таких директоров, работающих в восьми крупных компаниях по страхованию жизни. Из этих 80 человек 42, считавшихся руководством компании весьма успешными работниками, составили верхнюю критериальную группу. Остальные 38 человек, считавшиеся неудовлетворительными работниками, образовали нижнюю критериальную группу. Полученные 80 протоколов ответов были изучены экспертами по тесту Роршаха, отобравшими 32 признака (или характеристики ответов), чаще встречавшихся в одной группе, нежели в другой. Признаки, чаще обнаруживаемые в верхней критериальной группе, оценивались в +1 балл при их наличии и в 0 баллов при их отсутствии у обследуемого; признаки, чаще встречавшиеся в нижней критериальной группе, соответственно оценивались в −1 балл при их наличии и в 0 баллов при их отсутствии. Поскольку всего имелось по 16 признаков каждого типа, суммарный показатель теоретически мог принимать значения от −16 до +16.

Когда оценочный ключ, основанный на этих 32 признаках, был применен к первоначальной группе из 80 человек, принадлежность 79 из них к верхней или нижней группе была определена правильно. Таким образом, корреляция между тестовым показателем и критерием оказалась близкой к 1,00. Однако когда была проведена перекрестная валидизация теста на второй сопоставимой выборке коммерческих директоров страховых агентств, насчитывавшей 41 человек (21 в верхней и 20 в нижней критериальной группе), коэффициент валидности упал до пренебрежимо малой величины 0,02. Очевидно, таким образом, что ключ, разработанный на первой выборке, не был валидным, а значит, и пригодным, для отбора кандидатов на такую должность.

Пример со случайными данными. В классическом исследовании Кьюретона (Cureton, 1950) было получено яркое доказательство того, что при использовании одной и той же выборки для отбора заданий и валидизации теста можно получить полностью фиктивный коэффициент валидности даже при чисто случайных условиях. В этой работе прогнозируемым критерием служил средний балл каждого из 29 студентов, записавшихся на курс психологии. Весь диапазон значений этого критерия был разбит на две области: оценки не ниже «B» и оценки ниже «B». Роль «заданий» в этом эксперименте играли 85 номерков с числами от 1 до 85 на одной стороне. Чтобы получить тестовый показатель для каждого студента, номерки складывались в коробку, перемешивались и высыпались на стол. Те из них, которые падали лицевой стороной кверху, регистрировались как номера выполненных данным студентом заданий. Совокупный показатель каждого студента складывался из результатов 29 бросаний 85 номерков. Эту процедуру порождения случайных оценок Кьюретон в шутку назвал «тестом *B*-проективного психокинеза».

Затем был проведен анализ заданий, в котором в качестве критерия фигурировал средний балл студента. На этом основании из 85 «заданий» было отобрано 24, из которых 9 чаще встречались у студентов верхней критериальной группы и поэтому получили веса +1, тогда как 15 чаще выпадали в нижней критериальной группе, и им приписывались веса −1. Сумма весов «заданий» составляла суммарный тестовый балл каждого студента. Несмотря на заведомо случайное происхождение этих «тестовых баллов», их корреляция с критерием успеваемости для все той же группы из 29 студентов оказалась равной 0,82. Этот результат аналогичен тому, который был получен в примере с тестом Роршаха. В обоих случаях видимое соответствие между показателями теста и критерием вызвано использованием одних и тех же случайных различий как при отборе заданий, так и при определении валидности теста в целом.

Условия, влияющие на уменьшение валидности. Степень уменьшения коэффициента валидности при перекрестной валидизации частично зависит от размера первоначальной совокупности заданий и от того, какая часть заданий сохраняется. Если первоначальное число заданий велико, а доля отобранных заданий мала, то возрастает возможность использования случайных различий и тем самым получения искусственно завышенного коэффициента валидности. На степень уменьшения валидности при перекрестной валидизации влияет также объем выборки. Поскольку завышение валидности в первоначальной выборке является результатом накопления ошибок выборки, при малых выборках (для которых такие ошибки больше) будет наблюдаться большее снижение валидности.

Если задания отбираются на основе предварительно сформулированных гипотез, выводимых из психологической теории или опыта работы с данным критерием, то уменьшение валидности при перекрестной валидизации будет минимальным. Например, если согласно конкретной гипотезе ответ «да» должен появляться чаще среди успевающих учеников, то задание следует *отбросить*, когда ответ «да» значительно чаще исходит от *неуспевающих* учеников. Наоборот, полностью эмпирический подход означал бы включение в первоначальную совокупность самых разнообразных вопросов, безотносительно к их связи с критериальным поведением, в расчете на последующий отбор заданий, имеющих значимую положительную или отрицательную корреляцию с критерием. В последнем случае следует ожидать большего снижения валидности, чем в первом. В своем хорошо спланированном исследовании Митчелл и

Климоски (T. W. Mitchell, & Klimoski, 1986) убедительно продемонстрировали различия в уменьшении валидности, которое фактически имеет место при отборе заданий на основе рационального или эмпирического подхода. Итак, уменьшение валидности теста при перекрестной валидизации будет наибольшим, если выборки малы, исходная совокупность заданий велика, а доля отобранных из нее заданий мала, и если задания подбираются без заранее сформулированного рационального основания.

Дифференцированное функционирование заданий

Статистические процедуры. В качестве одного из аспектов исследования необъективности тестов в отношении групп меньшинств все большее внимание привлекает анализ «систематической ошибки задания» (*item bias*). Предметом такого анализа является, по существу, ***относительная трудность*** отдельных заданий теста для групп, различающихся культурными истоками и жизненным опытом. В психометрической терминологии эта область анализа заданий известна под названием дифференцированного функционирования заданий (сокращенно *DIF* — по первым буквам *differential item functioning*). Цель анализа *DIF* — идентифицировать задания, в отношении которых одинаково способные лица из различных культурных групп имеют разные вероятности успеха. Он основан на предположении, что одинаковая способность означает равенство в отношении конструкта, для оценки которого предназначен данный тест, или критериального поведения, для предсказания которого этот тест используется. Для идентификации таких дифференцированно функционирующих заданий было разработано множество методов, включая статистические и оценочные процедуры (Berk, 1982; Camilli, & Shepard, 1994; Hambleton, & Rogers, 1989; P. W. Holland, & Thayer, 1988; P. W. Holland, & Wainer, 1993; Osterlind, 1983; C. R. Reynolds, & Brown, 1984).

Главная проблема заключается в том, что демографические (или другие связанные с жизненным опытом) различия групп в трудности задания тесно связаны со среднегрупповыми различиями в уровне выполнения теста в целом. В результате, задания, обладающие хорошей различительной способностью с точки зрения суммарного показателя, могут выглядеть «необъективными» и, вследствие этого, отбрасываться. Для контроля за такими различиями в суммарном показателе использовалось несколько процедур. С расширением доступа к компьютерам одним из самых многообещающих становится метод, основанный на теории «задание — ответ» (*IRT*). Этот класс процедур особенно уместно применять в тех случаях, когда в распоряжении исследователей оказываются большие выборки. Как уже было показано в этой главе, характеристические кривые (*ICC*) для каждого задания показывают вероятность правильного ответа относительно шкалы способности теста (рис. 7–6). Сравнивая *ICC* для одного и того же задания в любых двух группах, мы можем идентифицировать задания со значимым дифференцированным функционированием относительно полного выполнения теста группами, выраженного в единой шкале. Рис. 7–7 иллюстрирует существо такого сравнения на примере двух заданий. Как легко заметить, для задания 1 характеристические кривые в группах *A* и *B* существенно различаются, тогда как для задания 2 они очень похожи. Для каждого задания область между двумя *ICC* можно использовать, чтобы установить диапазон способности, в котором содержатся признаки *DIF*. После того как *DIF* задания идентифицированы, какая бы процедура для этого ни

использовалась, следующий шаг — выяснение характера и источника установленного различия. Ответ на этот вопрос определяет, войдет ли оно в состав теста или будет отброшено. Для этой цели могут потребоваться различные оценочные процедуры (*judgmental procedures*), возможно в сочетании с последующим статистическим анализом.

Оценочные процедуры. Не существует какого-то одного, «наилучшего метода» анализа заданий, подходящего для всех целей. Поскольку разные методы дают в чем-то различные виды информации, желательно использовать их комбинацию. Целесообразное сочетание методов зависит от предполагаемого назначения теста и от характера выводов, делаемых из его показателей. Обычно, наилучшим оказывается некоторое сочетание статистических и оценочных процедур.

При правильном применении оценочные процедуры могут снабжать нас полезной информацией, которую невозможно получить иным способом (Scheuneman, 1982; Tittle, 1982). Анализ субъективных оценок особенно полезен на начальном и заключительном этапах конструирования теста, предваряя и завершая статистический анализ. На начальном этапе разработки теста оценочный анализ обычно проводится для того, чтобы отсеять содержание, которое может оскорблять или унижать меньшинства, либо укреплять социальные стереотипы в отношении профессиональных или других социальных ролей. С этой целью крупные издательства тестов регулярно практикуют предварительный просмотр заданий, привлекая к этому как своих сотрудников, так и консультантов со стороны, представляющих разные социокультурные группы (Berk, 1982, chap. 9). Такой просмотр также помогает выявить содержание теста, которое может ограничиваться рамками определенной культуры и потому быть незнакомым для отдельных популяций тестируемых. Следует, однако, заметить, что такие оценочные просмотры, как правило, *не* дают хороших результатов при предсказании относительной трудности или различительной способности заданий для различных популяций (Plake, 1980; Sandoval, & Miille, 1980; Scheuneman, 1982). Для этой цели необходим статистический анализ эмпирических результатов.

С другой стороны, далеко не все отклоняющиеся от нормы задания, выявленные с помощью статистических процедур, можно расценивать как необъективные. Результаты статистического анализа требуют интерпретации на основе второго просмотра заданий и совершенно иного рода оценочного анализа. На этой стадии задания изучаются на предмет возможных источников их статистической аномальности. Статистические выбросы не обязательно обнаруживают какую-то общую характеристику или явную причину отклонения; каждое задание требует индивидуального рассмотрения. Отдельные выбросы могут просто отражать статистические артефакты, возникающие в результате применения конкретной процедуры. В других случаях отклоняющееся выполнение задания может быть следствием любого из широкого множества условий, которые имеют различные следствия для интерпретации теста. Правильная оценка таких аномальных заданий требует знания как содержательной области теста, так и различий в опыте тестируемых, относящихся к разным популяциям.

Возможная причина аномальности заключается в том, что задание не измеряет один и тот же конструкт в разных группах. Например, словесные аналогии могут измерять вербальное рассуждение в одной группе и знание слов в другой, если такое задание содержит ключевое слово, незнакомое многим членам определенного меньшинства. Подобным же образом арифметическая задача может измерять математическую способность в одной группе и способность понимать сложные словесные формули-

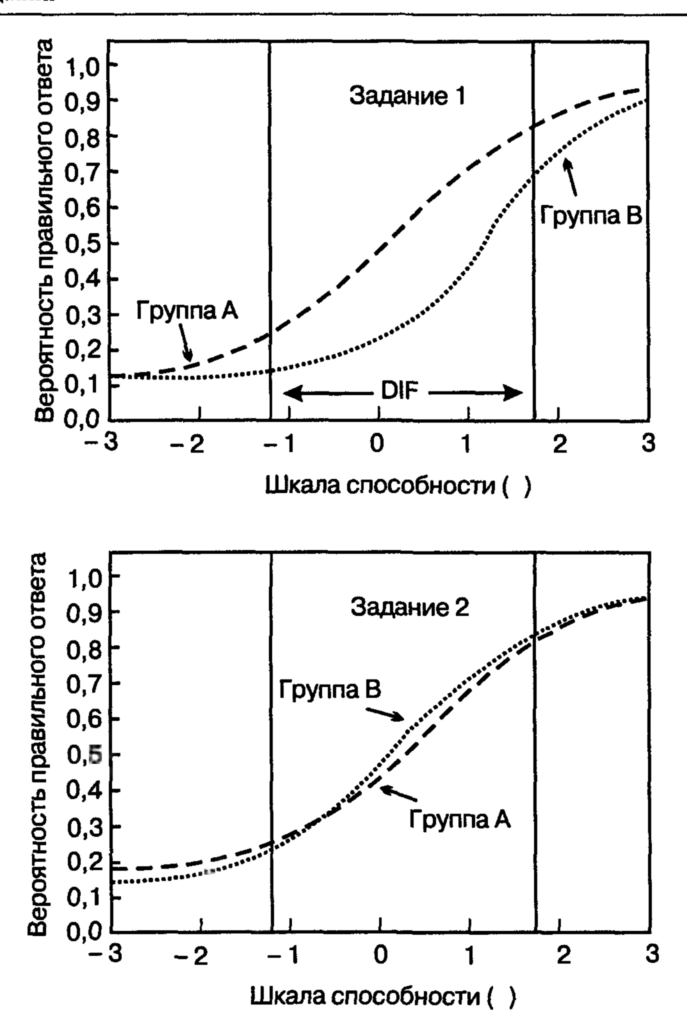


Рис. 7–7. Характеристические кривые (*ICC*) для двух заданий, иллюстрирующие разную степень дифференцированного функционирования задания (*DIF*)

(Графики, с некоторыми упрощениями, взяты из Pashley, 1992. Воспроизведено с разрешения)

ровки в другой. В этих двух примерах недостающие знание не имеет отношения к конструкту, измеряемому тестом в целом. Предположим, однако, что математические задания, включающие десятичные дроби, оказались относительно более трудными для членов какой-то конкретной группы. Это различие релевантно конструкту математической способности. Следовательно, такие выбросы не являются, в этом смысле, необъективными заданиями.

В тех случаях, когда аномальные задания идентифицируются статистически, источник этой аномальности можно отчасти прояснить, применяя дополнительные статистические процедуры, такие как анализ ошибочных вариантов ответа, выбираемых в задании со множественным выбором. Этот дополнительный анализ заданий, в сочетании с их критическим просмотром и оценкой, должен дать основание для соответствующего действия. Аномальное задание может быть отброшено, пересмотрено це-

ликом или частично изменено в его некорректной части; возможно, придется расширить или сделать более ясными инструкции к тесту, а может быть, задание будет сохранено в первоначальном виде после его повторного рассмотрения в свете спецификации теста. Анализ задания может даже потребовать переоценки самой этой спецификации, что ведет либо к ее изменению, либо к уточнению допустимых выводов из тестовых показателей.

Известный случай неправильного использования *DIF*. Широкую огласку получило дело, при рассмотрении которого суд, вероятно, впервые основывался главным образом на анализе заданий в оценке «необъективности теста». Этот прецедент стал известен общественности под названием «дело “Золотого правила”», поскольку оно было связано с проведением экзамена на получение лицензии при найме служащих страховой компанией *Golden Rule* («Золотое правило»). Сходство между названием компании и общим употреблением этих двух слов в совершенно ином смысле,[1] по-видимому, еще больше осложнило ситуацию. Окончательное решение по данному делу основывалось исключительно на сравнении групповых различий в проценте лиц, справившихся с заданием, без каких-либо попыток установить равенство групп по любому показателю способности, для оценки которой был предназначен тест, или рассмотреть валидность заданий относительно предполагаемой цели теста. Данное решение явно противоречило понятию дифференцированного функционирования задания и, по всей видимости, вело к исключению из теста тех самых заданий, которые были наилучшими предикторами выполнения работы.

Судебной ошибке, содержащейся в решении по делу «Золотого правила», вследствие ее непонимания широкой общественностью и возможного влияния созданного прецедента на использование тестов в профотборе и образовании, были даны критические оценки с разных сторон (например, Lim, & Drasgow, 1990), включая официальное заявление Американской психологической ассоциации.[2] Это судебное решение также стало темой симпозиума на ежегодном съезде АПА, большая часть докладов на котором впоследствии была опубликована в специальном выпуске журнала *Educational and Psychological Measurement: Issues and Practices* (Bond, 1987; Faggen, 1987; Linn, & Drasgow, 1987). Рассмотрение этого печально известного судебного случая высвечивает потенциальные практические опасности для тех, кто пытается оценивать «необъективность теста» по поверхностным и неполным признакам.

Поисковые исследования в области разработки заданий

Быстрое расширение использования компьютеров в 1980-е и 1990-е гг., в сочетании с достижениями когнитивной психологии, стимулировало широкие исследования в рамках новых подходов к разработке заданий. Традиционно составление заданий было скорее искусством, чем наукой. Даже при идеальных условиях составите-

[1] «Золотым правилом» принято называть библейскую заповедь: «Во всем, как хотите, чтобы другие поступали с вами, поступайте и вы с ними». — *Примеч. науч. ред.*

[2] Подготовленное Комитетом по психологическим тестам и психологическому оцениванию АПА, это заявление было одобрено соответствующими отделениями АПА и Советом представителей.

лям заданий давались инструкции, которые определяли лишь форму задания и охватываемое им содержание. Все еще распространена практика, когда разработчики опираются на предварительную эмпирическую проверку заданий, чтобы оценить их уровень трудности и различительную способность. Есть ли какой-то способ предсказать эти статистики задания до его предварительной проверки, только на основе анализа физических или семантических свойств стимулов? Или, что еще лучше, можно ли конструировать задания с требуемым уровнем трудности и различительной силы? Может ли систематическое манипулирование характеристиками стимула предопределять востребование заданиями теста определенных когнитивных процессов? Все это вопросы, исследование которых непрерывно ведется как экспериментальными, так и математическими методами (Bejar, 1985, 1991; Carroll, 1987; Embretson, 1985a, 1985b, 1991, 1994, 1995; Freedle, 1990).

Востребование, или запрос тестовыми стимулами определенных когнитивных процессов может исследоваться посредством методик декомпозиции задачи, разработанных в когнитивной психологии. Они позволяют устанавливать связи различных свойств задания со скоростью его выполнения и допускаемыми ошибками. Несколько таких исследований было проведено с пространственными заданиями (Embretson, 1994; Pellegrino, Mumaw, & Shute, 1985). Например, предъявляемые в тесте пространственных аналогий стимулы можно классифицировать относительно: 1) сложности, или количества отдельных элементов, которые должны быть распознаны (например, форма, размер, положение), и 2) преобразований, или числа способов, какими стимул изменяется в подлежащей оцениванию паре. В некоторых типах задач на пространственное воображение, требующих от тестируемого выбрать определенные части, из которых можно сложить заданную целую фигуру, эти части могут быть просто разнесенными в пространстве, смещенными, повернутыми или измененными сочетанием этих способов.

Предметом других исследований были семантические характеристики вербальных стимулов. Например, в тестах вербального рассуждения задания могут конструироваться в соответствии с известными логическими принципами и законами (Colberg, 1985; Colberg, Nester, & Trattner, 1985; Scheuneman, Geritz, & Embretson, 1991; K. Sheehan, & Mislevy, 1989; Shye, 1988). Такая процедура могла бы гарантировать, что только один из вариантов ответа является в подлинном смысле правильным и что различные логические отношения представлены в выборке заданий в заранее заданной пропорции. Кроме того, эта процедура дала бы возможность манипулировать логической сложностью задания, связь которой с его уровнем трудности можно было бы затем исследовать эмпирически. Некоторые исследователи экспериментировали с конструированием символических (в частности, буквенных) последовательностей, предназначенных для тестирования индуктивного рассуждения (Butterfield et al., 1985). Сначала был разработан полный набор правил для систематического конструирования таких последовательностей. Затем были сформулированы гипотезы в отношении операций, выполняемых людьми, пытающимися понять заложенные в них закономерности. Наконец, эти гипотезы проверяли в ходе эмпирических исследований трудности заданий на завершение последовательностей.

Эмбретсон (Embretson, 1994) предлагает радикальные изменения в анализе заданий и усовершенствование процесса их разработки. Весь процесс начинается с определения подлежащих оценке конструктов, после чего строится когнитивная модель для конструируемого теста. Подробные характеристики этой когнитивной модели обес-

печивают спецификацию для создания заданий. Затем проводится эмпирическая валидизация заданий, чтобы установить их фактическое соответствие теоретической когнитивной модели в ее практических приложениях. Полная процедура иллюстрируется разработкой Обучающего теста пространственной способности (*Spatial Learning Ability Test*), который измеряет не только исходный уровень пространственной способности, но и ее видоизменяемость после стандартизованного обучения.

Исследования предсказания трудности задания по физическим и семантическим свойствам стимулов не только помогает разработчикам заданий создавать эффективные тесты, но и подводит к автоматизированному, компьютерному конструированию заданий. Разумеется, детальную спецификацию задания можно без особого труда включить в машинную программу (см., например, Butterfield et al., 1985; Embretson, 1994). Бесспорно и то, что потенциальные преимущества этих развивающихся методов конструирования тестов впечатляют. И все же не следует ожидать слишком много от какого-то одного, пусть самого современного, подхода. Например, весьма вероятно, что тест может полно и эффективно измерять ряд четко идентифицированных конструктов и тем не менее не обладать высокой прогностической валидностью в некоторых важных областях его предполагаемого использования. По этой причине необходимо учитывать оба аспекта валидизации конструкта, которые Эмбретсон (Embretson, 1983) обозначает как репрезентацию конструкта и номотетический диапазон. Декомпозиция задачи дает информацию о репрезентации конструкта; определение номотетического диапазона требует изучения связей тестовых показателей в сети других, внешних переменных, включая и меры критерия. Другое предостережение против чрезмерной универсализации относится к необходимости обладать знанием релевантного содержания для эффективного выполнения задач в любой предметной области или сфере мастерства. Способы обработки информации часто связаны с содержанием, и потому не могут эффективно оцениваться в отсутствие соответствующего содержания.

В заключение отметим, что упоминавшиеся в этом разделе новаторские методы, при их правильном применении, могут внести существенный вклад в систематическое и управляемое конструирование тестовых заданий. Более того, благодаря идентификации измеряемых тестом конструктов, эти методы могут значительно улучшить наше понимание причин того, почему конкретные тесты предсказывают выполнение в критериальных ситуациях. Дополнительное преимущество касается диагностического использования тестов, поскольку источник сильных и слабых сторон индивидуума можно в этом случае связать с конкретными когнитивными процессами. Все это достойные цели, однако их практическая реализация еще требует значительных исследований оставшихся нерешенными проблем (см., например, Wainer, 1993 a). В настоящее время ведется большая исследовательская работа в области разработки заданий, допускающих идентификацию когнитивных процессов отдельных респондентов при решении конкретных задач (Willson, 1994). Анализ типов ошибок, совершаемых испытуемыми, открывает многообещающие пути к достижению этой цели (Kulikowich, & Alexander, 1994).

Часть 3

ТЕСТИРОВАНИЕ СПОСОБНОСТЕЙ

8 ИНДИВИДУАЛЬНЫЕ ТЕСТЫ

Во второй части мы познакомились с основными принципами психологического тестирования и теперь можем применить их для оценки конкретных тестов. Мы уже знаем, какие вопросы задать по поводу каждого теста и где искать на них ответы. Руководства по тестам и *Ежегодники психических измерений* (*Mental Measurements Yearbooks*) входят в число главных источников, к которым можно обратиться за получением информации в отношении любого из упоминаемых здесь тестов.[1]

Оставшиеся части книги преследуют двоякую цель. Во-первых, они предоставляют возможность проследить за применением принципов тестирования в широком множестве тестов. Во-вторых, познакомить читателя с некоторыми из наиболее характерных тестов в каждой из основных областей их применения, не пытаясь при этом дать их исчерпывающий обзор. Такой обзор не составляет цели данной работы и скорее всего устарел бы еще до выхода книги в свет из-за той быстроты, с какой появляются новые тесты или их пересмотренные версии. По этим причинам в каждой разновидности тестов обсуждаются лишь несколько наиболее типичных, выбранных либо из-за их общеупотребительности, либо из-за того, что они иллюстрируют важные достижения в процедуре тестирования. При этом тестирование способностей рассматривается в части 3, тестирование личности — в части 4 и применение тестирования в разных средах, или контекстах — в части 5. Если не оговорено особо, следует иметь в виду, что все данные об обсуждаемых в этой книге тестах берутся из руководств по конкретным тестам или специальных приложений, которыми издатели снабжают те или иные тесты. Читатели, желающие самостоятельно провести критический разбор какого-то конкретного теста, могут воспользоваться схемой оценки теста, предложенной в *Study Guide* к этому учебнику (Urbina, 1997). Более подробные указания для этого даны в *Стандартах тестирования (Testing Standard)* (AERA, APA, NCME, 1985).

Обсуждаемые в этой и следующей главах виды тестов, традиционно называемые «тестами интеллекта», ведут свое происхождение от шкал Бине. Такие тесты предназначены для использования в достаточно разнообразных ситуациях, а их валидность

[1] Десятитомная серия *Test Critiques* (Keyser & Sweetland, 1984–1994) служит другим полезным источником информации и критических оценок в отношении сотен тестов.

устанавливается с применением относительно широких критериев (см. L. R. Aiken, 1996). Как правило, они дают один суммарный показатель, такой как традиционный *IQ* или индекс общего уровня выполнения теста обследуемым. Кроме того, они обычно дают показатели по отдельным субтестам или их группам, оценивающие более узко определяемые способности (*aptitudes*). Поскольку валидность большей части тестов интеллекта устанавливалась относительно мер учебных достижений, их часто называют тестами академических способностей или академического интеллекта. Тесты интеллекта нередко используют в качестве инструментов предварительного отсеивания, после которого уже с меньшим числом кандидатов проводят тесты специальных способностей. Такая практика особенно распространена в тестировании нормальных подростков и взрослых при консультировании по вопросам обучения или выбора профессии, подборе кадров и решении других схожих задач. Еще одной областью широкого применения тестов общего интеллекта является клиническое тестирование, особенно в той его части, которая касается распознавания и классификации лиц с умственной отсталостью. Для этих целей обычно используют индивидуальные тесты, среди которых наиболее употребительными (в противопоставлении групповым) можно назвать обсуждаемые в этой главе шкалы Стэнфорд—Бине и Векслера. Поскольку шкала Стэнфорд—Бине — это первый тест, освещаемый в данной книге, он рассматривается полнее других тестов, обсуждаемых на всем протяжении учебника. Это сделано для того, чтобы с самого начала проиллюстрировать все виды информации, принимаемой в расчет при оценивании теста. Следует, однако, отметить, что обсуждения конкретных тестов на страницах этой книги не нужно рассматривать как их критические обзоры, подобные тем, которые даются, например, в *Ежегодниках психических измерений*. В соответствии с целями нашего учебника предметом внимания обычно становятся особые достоинства конкретного теста или характерные особенности, отличающие его от других тестов.[1]

Шкала интеллекта Стэнфорд—Бине

Развитие шкал интеллекта. Исходные шкалы Бине—Симона, опубликованные во Франции в 1905, 1908 и 1911 гг., вкратце уже были охарактеризованы в главе 2. Напомним только, что среди многочисленных переводов и адаптаций ранних тестов Бине, появившихся в США, самым жизнеспособным оказался тест Стэнфорд—Бине.[2] Первая стэнфордская редакция шкал Бине—Симона, подготовленная Л. М. Тёрменом и его коллегами в Стэнфордском университете, была опубликована в 1916 г. (Terman, 1916). В ней было введено так много изменений и дополнений, что фактически она представляла собой новый тест. Более трети заданий были заменены новыми, а ряд старых или переделан, или перераспределен по другим возрастным уровням, или отброшен. Вся шкала была заново стандартизована на национальной выборке, состоявшей приблизительно из 1000 детей и 400 взрослых. Были подготовлены подробные инструкции по проведению теста и подсчету баллов, и впервые был использован пока-

[1] Отличный обзор многих тем, обсуждаемых в части 3 учебника, можно найти в книге *Contemporary Intellectual Assessment* (Flanagan, Genshaft, & Harrison (Eds.), 1997).

[2] Подробный разбор шкал Бине—Симона и сводку данных о развитии, использовании и клинической интерпретации шкал Стэнфорд—Бине можно найти у Sattler (1982, 1988).

затель *IQ*,. Вторая стэнфордская редакция теста, появившаяся в 1937 г., состояла из двух эквивалентных форм *L* и *M* (Terman, & Merrill, 1937). В этом варианте шкала была значительно увеличена в объеме и полностью рестандартизована на новой выборке населения США. Однако несмотря на все усилия получить срез, адекватно представляющий структуру населения, выборка из 3184 обследованных оказалась несколько выше по социально-экономическому уровню, чем все население США, содержала избыток городских жителей и включала только представителей коренного белого населения.

Опубликованная в 1960 г. третья редакция предусматривала единственную форму (*L–M*), объединившую в себе лучшие задания двух форм 1937 г. (Terman, & Merrill, 1960). При подготовке шкалы Стэнфорд—Бине 1960 г. ее авторы столкнулись с общей дилеммой психологического тестирования. С одной стороны, частые переделки теста желательны, поскольку позволяют воспользоваться новыми наработками в конструировании тестов и накопленным опытом применения теста, а также постоянно обновлять содержание теста. Последнее особенно важно для заданий на осведомленность и для используемого в тесте наглядного материала, содержание которого подвержено влиянию моды: изменению фасонов одежды, домашней утвари, машин и других бытовых предметов. Использование теста с устаревшим содержанием может серьезно нарушить раппорт между тестируемым и тестирующим и повлиять на уровень трудности заданий. С другой стороны, пересмотры теста могут привести к тому, что значительная часть накопленных данных окажется неприменимой к его новой форме. По тестам, широко применявшимся многие годы, накапливается большой материал по интерпретации их результатов, значимость которого необходимо тщательно взвесить относительно потребности в пересмотре теста. По этим соображениям создатели шкалы Стэнфорд—Бине предпочли свести две прежние формы в одну, выбирая тем самым золотую середину между опасностью устаревания и нарушения преемственности теста. Утрата параллельной формы не была слишком большой платой за достижение этой цели. В 1960 г. необходимость во взаимозаменяемой форме ощущалась менее остро, чем в 1937 г., когда не существовало иных достаточно надежных индивидуальных шкал интеллекта. Редакция Стэнфорд—Бине 1960 г. *не* предусматривала рестандартизации нормативной шкалы. Новые выборки были использованы только для того, чтобы выявить изменения в трудности заданий, происшедшие за истекший период. В результате, показатели умственного возраста и *IQ* в форме *L–M* 1960 г. по-прежнему выражались на основе нормативной выборки 1937 г.

Следующей стадией была рестандартизация формы *L–M*, проведенная в 1972 г. (Terman, & Merrill, 1973, Pt. 4). На этот раз содержание теста осталось практически неизменным, а нормы были получены на новой выборке, состоявшей приблизительно из 2100 человек, протестированных в 1971/72 учебном году. По сравнению с нормами 1937 г. нормы 1972 г. основывались на более репрезентативной выборке и, будучи более современными, отражали влияние происшедших за это время культурных перемен на выполнение теста. Интересно отметить, что эти нормы показали некоторое улучшение в выполнении теста во всех возрастных группах. Существенное улучшение наблюдалось в дошкольном возрасте, в среднем на 10 единиц *IQ*. Авторы теста относят это улучшение на счет воздействия на маленьких детей средств массовой информации, роста грамотности и общего образовательного уровня родителей, равно как и других изменений в культуре. Наблюдалось также несколько меньшее, но заметное повышение уровня выполнения теста в возрасте 15 лет и старше, что, как полагают

авторы, может быть связано с увеличением в 1970-х гг. (по сравнению с 1930-ми) доли учащихся, продолжающих свое образование в средней школе до конца. На основе сравнения данных, полученных как методом поперечных срезов, так и в лонгитюдных исследованиях, Р. Л. Торндайк (R. L. Thorndike, 1977) изучил эти изменения норм в более широкой вре́менной перспективе и высказал предположение о действии ряда других факторов, включая введение специальных телепрограмм для стимулирования интеллектуального развития детей дошкольного возраста.

Повышение тестовых норм в период с 1930-х или 1940-х гг. по 1970-е гг. было обнаружено и в других тестах, используемых для оценки общего интеллектуального уровня (Flynn, 1984, 1987). С точки зрения пользователя теста важным следствием из таких данных будет то, что отдельные люди или группы, обследуемые с помощью ранних и поздних тестовых форм, обнаружат снижение способности, поскольку выполнение ими теста оценивается относительно более высокого стандарта поздней формы. Проводящий обследование должен иметь в виду этот возможный артефакт при интерпретации показателей.

Четвертая редакция шкалы Стэнфорд—Бине (SB-IV): Общая характеристика. Современная редакция этой хорошо зарекомендовавшей себя шкалы является результатом наиболее обширного ее пересмотра (Delaney, & Hopkins, 1987; Thorndike, Hagen, & Sattler, 1986a, 1986b). Сохраняя главные преимущества более ранних редакций как индивидуально применяемого клинического инструмента, эта версия отражает результаты развития как теоретических представлений об интеллектуальных функциях, так и методологии конструирования тестов. Преемственность с более ранними редакциями была отчасти обеспечена путем сохранения многих типов заданий из ранних форм. Еще важнее, что удалось сохранить адаптивную процедуру тестирования, благодаря которой каждый тестируемый получает только те задания, чья трудность соответствует продемонстрированному им уровню выполнения.

В то же время сфера содержания была сильно расширена по сравнению с преимущественно вербальным фокусом ранних форм, с тем чтобы обеспечить более репрезентативный охват задач на оперирование числами, пространственными отношениями и данными кратковременной памяти. Кроме того, каждый тип заданий используется, насколько это возможно, в широком возрастном диапазоне, обеспечивая тем самым почти полную сопоставимость оценок на разных возрастных уровнях. Четвертая редакция шкалы Стэнфорд—Бине предназначена для использования в возрастном диапазоне от двух лет до взрослости.

Проведение тестирования и подсчет баллов. Типовой набор материалов, необходимых для проведения теста Стэнфорд—Бине, показан на рис. 8–1. В него входят четыре книжечки отпечатанных типографским способом карточек с изображениями тестовых заданий, смена которых осуществляется перебрасыванием страниц; предметный материал теста, включающий кубики, доску (геометрических) форм, набор разноцветных и имеющих разную форму бусинок, а также большую картинку с изображением неразличимой по полу и этническим признакам куклы; тетрадь с протоколами для регистрации ответов и руководство по проведению теста и оценки результатов.

Как и большинство индивидуальных тестов интеллекта, шкала Стэнфорд—Бине требует, чтобы с ней работали только высококвалифицированные специалисты. Специальная подготовка и опыт работы с этой шкалой совершенно необходимы для пра-



Рис. 8–1. Материалы, используемые при проведении тестирования с помощью шкалы интеллекта Стэнфорд—Бине (четвертая редакция)
(Copyright © 1986 by the Riverside Publishing Company. Воспроизведено с разрешения издателя)

вильного проведения, подсчета баллов и интерпретации результатов теста. Неуверенность и неумелость могут губительно сказаться на раппорте, особенно с маленькими детьми. Незначительные изменения в словесных формулировках, допускаемые по невнимательности, могут изменить трудность заданий. Дополнительные сложности возникают в связи с тем, что задания должны оцениваться сразу же после их выполнения, поскольку последующее проведение испытания зависит от того, как обследуемый справился с заданиями предыдущих уровней.

Десятилетиями клиницисты относились к шкале Стэнфорд—Бине и подобным ей индивидуальным шкалам не только как к набору стандартизованных тестов, но и как к клиническому интервью. Те же особенности, которые затрудняют применение таких шкал, создают благоприятные возможности для взаимодействия диагноста и обследуемого и позволяют опытному клиницисту выявить необходимую ему для диагноза информацию. Шкала Стэнфорд—Бине и другие тесты, описанные в этой главе, позволяют наблюдать методы работы респондента, его подходы к решению задач и другие качественные аспекты выполнения заданий. Проводящий тестирование имеет также возможность оценить некоторые эмоциональные и мотивационные характеристики тестируемого, такие как способность сосредоточиться, уровень активности, уверенность в себе и настойчивость. Конечно, любые качественные наблюдения, делаемые в момент проведения индивидуальных тестов, необходимо фиксировать именно как наблюдения, а не интерпретировать тем же способом, что и объективные тестовые показатели. Ценность таких качественных наблюдений сильно зависит от мастерства, опыта и психологического чутья проводящего тестирование специалиста, равно как и от знания ловушек и ограничений, свойственных этому виду наблюдения.

Возраст 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18+

Вербальное рассуждение

Словарный

Понимание

Нелепости

Вербальные отношения

Количественное рассуждение

Количественный

Числовые ряды

Составление равенств

Абстрактное / наглядное рассуждение

Анализ конфигураций

Копирование

Матрицы

Складывание и разрезание бумаги

Кратковременная память

Память на бусинки

Память на фразы

Память на цифры

Память на предметы

Рис. 8–2. Возрастной диапазон 15 тестов шкалы Стэнфорд–Бине (четвертая редакция)

Примечание, касающееся областей, закрашенных серым цветом. Что касается девяти тестов с ограниченными возрастными диапазонами, некоторым членам выборки стандартизации, выходящим за их границы, все же предъявлялись какие-то из этих тестов из-за необычайно высокого или низкого результата по тесту, определяющему маршрут тестирования. Их показатели учитывались при оценивании результатов всей соответствующей возрастной выборки для составления нормативных таблиц, но эти оценки включались в них со специальным предостережением в отношении их использования. Что касается деталей, см. *Guide* (Thorndike et al., 1986a, p. 7) и *Technical Manual* (Thorndike et al., 1986b, p. 30).

(Приведено с упрощениями из The Stanford-Binet Intelligence Scale: Fourth Edition, Guide for administering and scoring, *p. 7. Copyright © 1986 by the Riverside Publishing Company. Воспроизведено с разрешения издателя)*

В отличие от возрастного принципа группировки заданий, применяемого в более ранних редакциях шкалы, в *SB*-IV задания каждого типа помещены в отдельные тесты в порядке возрастания трудности. Шкала состоит из 15 тестов, подобранных таким образом, чтобы представлять четыре основные когнитивные области: вербальное рассуждение, абстрактное/наглядное рассуждение, количественное рассуждение и кратковременную память (см. рис. 8–2). Эти 15 тестов, хотя и сгруппированы в четыре категории в целях вычисления показателей, проводятся в смешанном порядке для поддержания интереса и внимания тестируемых. Диапазон трудности шести из этих тестов перекрывает весь возрастной диапазон шкалы *SB*-IV. Как можно увидеть на

рис. 8–2, остальные девять тестов, вследствие характера содержащихся в них задач, либо начинают предъявлять позже, либо перестают предъявлять раньше соответствующих предельных возрастных уровней.

Проведение *SB*-IV представляет собой двустадийный процесс. На первой стадии тестирующий дает Словарный тест, который служит для выбора маршрута обследования через определение *начального уровня (entry level)* для всех остальных тестов. С какого задания начать Словарный тест зависит исключительно от хронологического возраста тестируемого. Для остальных тестов начальный уровень определяется по номограмме (или таблице) исходя из показателя Словарного теста и хронологического возраста. На второй стадии тестирования проводящий его специалист должен установить *базальный (basal)* и *предельный (ceiling) уровни* для каждого теста на основе фактического выполнения тестов индивидуумом. Базальный уровень достигается в том случае, когда испытуемый справляется с четырьмя заданиями на двух соседних уровнях. Предельный уровень достигается, когда три из четырех заданий (или все четыре задания) на двух соседних уровнях не выполняются испытуемым. По достижении предельного уровня по конкретному тесту его перестают использовать в дальнейшем тестировании испытуемого.

Когда задание предъявлено и на него получена реакция испытуемого, проводящий тестирование заносит оценку в тетрадь для записи ответов. Первичная оценка («сырой балл») по каждому тесту находится путем фиксирования номера задания самого высокого уровня из всех предъявленных испытуемому и вычитания из получившегося числа суммарного количества заданий, которые он выполнил неправильно. Кроме того, в состав 11 тестов входят задания-образцы, служащие лишь для ознакомления с тестом и никогда не учитываемые при вычислении показателя. В большинстве тестов каждое задание имеет только один верный ответ; такие ответы указаны на обратной стороне карточек с заданиями и в тетради для записи ответов. Все задания оценивают по принципу «выполнено/не выполнено», в соответствии с установленными эталонными ответами. Пять тестов предполагают свободные ответы, и потому требуют использования более развернутых нормативов и правил оценивания, которые даны в руководстве к проведению и оценке результатов *SB-IV* (Thorndike et al., 1986a),[1] где приведены и некоторые образцы двусмысленных ответов, требующих дополнительного уточнения со стороны проводящего тестирование специалиста.

Хотя полная шкала *SB*-IV имеет в своем составе 15 тестов, ни один человек не проходит все эти тесты, поскольку часть из них применима только в ограниченных возрастных диапазонах. Обычно полная батарея включает от 8 до 13 тестов, в зависимости от возраста тестируемого и его результата по тесту, определяющему маршрут обследования. Время проведения полной батареи предположительно колеблется от 30 до 90 минут, но менее опытным пользователям может потребоваться и больше времени. Как правило, обследование с помощью шкалы *SB*-IV проводится за один сеанс, возможно с перерывами в несколько минут между тестами. Для некоторых целей в руководстве по проведению и оценке результатов *SB*-IV (Thorndike et al., 1986a) предлагается несколько сокращенных батарей, требующих меньшего времени тестирования, но сфокусированных на тестах, наиболее подходящих для конкретной цели тестирования. В число таких батарей входят 6-тестовая сокращенная батарея общего

[1] К числу этих тестов относятся: Словарный, Понимание, Нелепости, Копирование и Вербальные отношения.

назначения и 4-тестовая батарея экспресс-скрининга. Обе имеют в своем составе по меньшей мере один тест в каждой из четырех когнитивных областей. Кроме того, предлагаются три батареи для обследования учащихся с целью включения в программы для одаренных детей, соответственно для каждого из трех возрастных уровней, и три батареи для учащихся с трудностями в обучении, также соответствующие трем возрастным уровням. Во всех этих сокращенных батареях используются стандартные процедуры для определения начальных уровней, проведения тестирования и подсчета баллов. В «Справочном руководстве для пользователей SB-IV» *(Examiner's Handbook)* (Delaney, & Hopkins, 1987) разъясняются многие процедурные вопросы, касающиеся проведения (и оценки результатов) этого теста с различными категориями обследуемых.

Стандартизация и нормы. Объем выборки стандартизации SB-IV немного превышал 5000 испытуемых в возрасте от 2 до 23 лет, протестированных в 47 штатах (включая Аляску и Гавайи) и округе Колумбия. Эта выборка была стратифицирована по таким признакам, как географический район, размер общины (*community size*), этническая группа и пол, с целью достичь близкого соответствия (на уровне пропорциональности) данным переписи населения США 1980 г. В добавление к этому контролировался социоэкономический статус испытуемых в виде профессионального и образовательного уровня родителей. Результаты этого контроля обнаружили избыточную представленность испытуемых на верхнем и недостаточную представленность на нижнем уровнях. Эти несоответствия были скорректированы путем приписывания различных весовых коэффициентов частотам при расчете значений показателя в нормативных таблицах. Таким образом, каждый испытуемый из семьи с высоким социоэкономическим статусом засчитывался как какая-то часть наблюдаемого случая, тогда как испытуемый из семьи с низким социоэкономическим статусом учитывался как случай с некой добавкой.

Нормативные таблицы используются для преобразования первичных показателей по каждому из 15 тестов в «стандартные показатели возраста» (Standard Age Scores, или, сокращенно, *SAS*).[1] Они представляют собой нормализованные стандартные показатели со средним, равным 50, и $SD = 8$ в каждой возрастной группе. Нормативные таблицы составлены с 4-месячным интервалом для возраста от 2 до 5 лет, с 6-месячным интервалом для возраста от 6 до 10 лет и с интервалом в 1 год для возраста от 11 до 17 лет; для возрастного уровня от 18 до 23 лет имеется одна-единственная нормативная таблица. Тетрадь для записи ответов содержит специальный бланк-диаграмму для построения индивидуального профиля *SAS* по результатам проведенных с конкретным испытуемым тестов.

Стандартные показатели возраста (*SAS*) можно также получить для каждой из четырех когнитивных областей и для совокупного результата по полной шкале *SB*-IV. Комплексный и четыре частных стандартных показателя возраста находят по значениям *SAS* для тестов, проведенных с конкретным испытуемым, для чего нужно просто обратиться к соответствующим нормативным таблицам. Эти пять *SAS* тоже являются

[1] Эти таблицы приведены в Thorndike et al., 1986a, p. 183–188. Некоторые значения *SAS*, основанные на менее 100 наблюдаемых случаев, статистически оценивались для полной возрастной когорты и выделены в нормативных таблицах темным фоном. Такие показатели появлялись тогда, когда испытуемые показывали необычайно высокий или, наоборот, низкий для своего возраста результат по тесту, определяющему маршрут обследования (Thorndike et al., 1986b, p. 29–30).

нормализованными стандартными показателями, но со средним, равным 100, и SD = 16. Таким образом, они выражаются в тех же единицах, что и стандартный *IQ* более ранних редакций шкалы Стэнфорд—Бине. Однако от использования термина «*IQ*» теперь полностью отказались. Для специальных целей предусмотрены возможности вычисления стандартных показателей возраста для любой комбинации двух или более частных (т. е. соответствующих одной из четырех когнитивных областей) *SAS* — так называемых «парциальных композиций» (*partial composites*). Например, комбинация *SAS* для вербального и количественного рассуждения близко соответствует «способности к обучению» (*scholastic aptitude*) и может представлять особый интерес в связи с оценкой академических достижений или готовности к обучению.

Надежность. Поскольку в *SB*-IV нет альтернативной формы, надежность этой шкалы можно было оценить только вычисляя внутреннюю согласованность или проводя повторное тестирование. В большинстве случаев использовался метод Кьюдера—Ричардсона, который применяли к данным, полученным на всей выборке стандартизации. Как и ожидалось, комплексный показатель по полной батарее дал наибольшие коэффициенты надежности на всех возрастных уровнях, значения которых колебались от 0,95 до 0,99. Надежность частных показателей в каждой из четырех когнитивных областей также оказалась высокой. Хотя она и изменялась в зависимости от числа тестов, включаемых в каждую область, соответствующие коэффициенты надежности варьировали в пределах от 0,80 до 0,97. Что касается отдельных тестов, то у большинства из них коэффициенты надежности попадают в интервал между 0,80 и 0,90, за исключением короткого (состоящего из 14 заданий) теста «Память на предметы», надежность которого варьирует от 0,66 до 0,78. В общем, все коэффициенты надежности имеют тенденцию несколько повышаться при переходе от младших к старшим возрастным уровням.

Дополнительные данные по ретестовой надежности были получены на 57 дошкольниках (5 лет) и 55 школьниках (8 лет), повторное тестирование которых проводилось спустя несколько месяцев (от 2 до 8). В общем, надежность оказалась высокой у комплексного показателя: соответствующие коэффициенты для этих двух групп составили 0,91 и 0,90. Хотя частный показатель в области вербального рассуждения дал коэффициенты надежности выше 0,80, ретестовая надежность других частных показателей и отдельных тестов обнаружила существенные колебания. Эти результаты трудно интерпретировать из-за возможного влияния ограниченных возрастных диапазонов некоторых тестов и эффекта практики, который мог существенно различаться от ребенка к ребенку.

В добавление к коэффициентам надежности в руководстве по проведению и оценке результатов *SB*-IV (*Guide*) и в техническом руководстве (*Technical Manual*) приводятся стандартные ошибки измерения (*SEM*) в пределах каждого возрастного уровня для каждого теста, частных показателей по когнитивным областям и комплексного показателя по полной шкале. Такие *SEM* нужны для оценивания индивидуальных показателей и для интерпретации различий между показателями при анализе профиля. Общий комплексный *SAS* ($M = 100$, $SD = 16$) имеет *SEM* от 2 до 3 единиц шкалы. Например, если в качестве приближенного среднего значения *SEM* взять 2,5, т. е. 2 шанса к 1, что «истинный» комплексный показатель конкретного испытуемого не будет отличаться от полученного им показателя больше чем на 2,5 единицы; кроме того, есть 95 шансов из 100, что его вариация составит не более 5 единиц ($2{,}5 \times 1{,}96 = 4{,}90$).

В *Справочном руководстве для пользователей SB-IV* (Delaney, & Hopkins, 1987) представлена интерпретационная основа, побуждающая формулировать гипотезы и проводить их перекрестную проверку на основе количественных и качественных данных, собранных с помощью этой батареи. Количественный анализ следует модели, впервые предложенной Ф. Б. Дэвисом (F. B. Davis, 1959) и примененной Кауфманом (Kaufman, 1979, 1994) и др. к шкалам Векслера. В сущности, он состоит из типовых схем сравнений комплексного и четырех частных (см. рис. 8–2) показателей с целью обнаружения статистически значимых различий исходя из величины *SEM*. Частоту полученных различий также сравнивают с соответствующими нормативными данными из выборки стандартизации. В дополнение к этому могут систематически оценивать сильные и слабые стороны конкретных способностей индивидуума, выявляемых каждым тестом, для чего проводят сравнения среднего результата испытуемого по комплексному и частным показателям с показателями по отдельным тестам. Указанное справочное руководство содержит всю необходимую информацию для проведения этих разновидностей анализа профиля, а также дает четыре полных примера их применения; оно наверняка будет оценено по достоинству как начинающими, так и опытными пользователями шкалы Стэнфорд—Бине.

Валидность. В соответствии с современными концепциями валидизации тестов разработчики четвертой редакции шкалы Стэнфорд—Бине придерживались разнообразных подходов при идентификации и определении закладываемых в ее основу конструктов. Первичный выбор конструктов направлялся результатами анализа доступной научной литературы о природе и измерении интеллекта (R. L. Thorndike et al., 1986b, chap. 1). Опыт использования прежних редакций этой шкалы и обнаружившиеся в ходе него ее сильные и слабые стороны служили дополнительными ориентирами при составлении планов конструирования новой шкалы и принятии решений. Например, разделение типов заданий на надежные субтесты было необходимой заменой традиционной клинической практики нестрогого анализа структуры ответов на основе субъективных группировок заданий.

После первичного выбора и предварительного определения конструктов, оцениваемых в *SB*-IV, были идентифицированы старые и разработаны новые задания, соответствующие этим определениям. Вся совокупность заданий подвергалась всестороннему и статистически изощренному анализу, включая как субъективную, так и статистическую оценку необъективности задания (R. L. Thorndike et al., 1986b, chap. 2). Окончательная версия шкалы, полученная в результате нескольких предварительных проверок и полевых испытаний, была проведена на выборке стандартизации и затем исследована в аспекте трех основных типов данных валидизации: 1) интеркорреляции и факторного анализа показателей; 2) корреляции с другими тестами интеллекта и 3) сравнения результатов в заранее установленных особых группах (Thorndike et al., 1986b, chap. 6).

Прежде всего, по данным полной выборки стандартизации вычисляли интеркорреляции между показателями всех тестов, частными показателями для четырех когнитивных областей и комплексными показателями батареи — отдельно по каждому возрастному уровню. Медианные корреляции (найденные ранжированием однотипных коэффициентов для всех возрастов) использовали в качестве исходных данных для конфирматорного (подтверждающего) факторного анализа. Главной целью этого анализа была проверка гипотезы о наличии общего фактора, объясняющего корреля-

ции между тестами из разных когнитивных областей, и групповых факторов, объясняющих остаточные корреляции внутри каждой области. Аналогичный факторный анализ также проводился с медианными корреляциями в каждой из трех возрастных групп (от 2 до 6, от 7 до 11 и от 12 до 18–23 лет).

Результаты факторного анализа в каждом случае показали существенные нагрузки общего фактора во всех тестах, оправдывая таким образом использование общего комплексного показателя. Для трех из четырех когнитивных областей групповые факторы объяснили значительную долю остаточной общей дисперсии внутри соответствующей области. Исключение составила область «абстрактного/наглядного рассуждения», где все четыре теста обнаружили высокую степень специфичности. Можно высказать предположение, что неспособность найти ясное подтверждение группового фактора в этой когнитивной области могла быть связана с кумулятивными эффектами школьного курса обучения, которое не так тщательно организовано в отношении пространственно-перцептивного содержания, как в отношении словесного и числового материала. Повседневный личный опыт, способствующий развитию пространственно-перцептивных способностей, не организуется систематически в «учебные курсы» или области содержания, подобно опыту, связанному с обучением. Поэтому менее вероятно, что личный опыт благоприятствует формированию общих структур связей у различных людей (Anastasi, 1970, 1986b).

Обзор результатов факторного анализа, приведенных в руководстве к тесту, так же как и результаты факторного анализа, проведенного независимо другими исследователями по данным стандартизации *SB*-IV, подтвердили правомерность использования комплексного показателя как меры общей интеллектуальной способности (R. M. Thorndike, 1990). Однако исследователи расходятся в том, что касается числа и природы более узких факторов (см. также McCallum, 1990). Эта ситуация осложняется тем, что поскольку *SB*-IV состоит из различных наборов тестов в разных возрастах, «сырые» данные для факторного анализа (т. е. корреляции между тестовыми показателями) различаются соответственно. Отсюда и различия в типах и количестве факторов — в пределах от двух до четырех, — появляющиеся на разных возрастных уровнях. Эти расхождения усугубляются разнообразием применяемых в разных исследованиях способов факторного анализа. Однако, в общем, с увеличением возраста испытуемых факторное решение лучше соответствует четырехфакторной модели, постулированной при разработке SB-IV, в особенности при использовании конфирматорного факторного анализа в противоположность эксплораторному (разведочному).

Второй источник данных валидизации основан на серии исследований групп, в которых проводился *SB*-IV и какой-нибудь другой тест интеллекта, включая форму *L–M* самой шкалы Стэнфорд–Бине.[1] Эти группы состояли из школьников, систематически посещающих занятия и охарактеризованных учителями как «обычные» (*nonexceptional*). Кроме того, в распоряжении исследователей были три «особые» (*exceptional*) группы учащихся, занимавшихся по программам для одаренных детей, детей с трудностями в обучении и детей с задержкой психического развития. В обычной выборке корреляция стандартного *IQ* по более ранней редакции шкалы Стэнфорд–Бине (форма *L–M*) с комплексным показателем по *SB*-IV составила 0,81; второй по величине (0,76) оказалась корреляция стандартного *IQ* формы *L–M* с частным пока-

[1] К числу других относились *WISC-R*, *WAIS-R*, *WPPSI* и *K-ABC*, которые будут рассмотрены в этой главе чуть позже.

зателем *SB*-IV в области «вербального рассуждения», а самую низкую корреляцию (0,56) стандартный *IQ* дал с частным показателем *SB*-IV в области «абстрактного/наглядного рассуждения», что и следовало ожидать исходя из сходства и различия в содержании этих двух форм шкалы Стэнфорд—Бине. Во всех группах корреляции комплексного и частных показателей *SB*-IV с общим или парциальными показателями по другим тестам интеллекта большей частью не противоречили гипотезам в отношении тестируемых конструктов. В то же время тщательное изучение всех корреляций, обнаруженных между специфическими показателями *SB*-IV и других тестов интеллекта способствует более твердому пониманию конструктов, измеряемых современной шкалой Стэнфорд—Бине.

Третья серия специальных исследований на особых выборках показала, что *SB*-IV позволяет правильно определять уровень выполнения одаренных, имеющих трудности в обучении и отстающих в развитии детей школьного возраста. Средние комплексного показателя и четырех частных показателей в выборке одаренных оказались существенно выше соответствующих средних в выборке стандартизации. Средние в выборках детей с трудностями в обучении и с задержкой психического развития были значимо ниже средних выборки стандартизации, а средние умственно отсталых — значимо ниже средних в выборке имеющих трудности в обучении. Следует заметить, что во всех исследованиях особых групп их участники определялись на основе тестов или других показателей деятельности, но сама шкала *SB*-IV при этом не использовалась.

В более позднем обзоре исследований валидности *SB*-IV (Laurent, Swerdlik, & Ryburn, 1992) делается вывод, что эта шкала является, по меньшей мере, столь же хорошим средством измерения общей интеллектуальной способности, как и другие имеющиеся в наличии средства; что она сильно коррелирует с мерами достижения и к тому же позволяет различать умственно отсталых, одаренных и больных с неврологическими повреждениями. Авторы обзора предполагают, что *SB*-IV можно использовать в качестве инструмента отбора при оценивании одаренных детей вследствие высокого «потолка», обеспечиваемого возрастным диапазоном этого теста; с другой стороны, они критикуют *SB*-IV за отсутствие предельно легких заданий — достаточно простых, чтобы диагностировать задержку умственного развития у самых маленьких детей.

Исследования, необходимые для усиления интерпретационного значения показателей различных тестов *SB*-IV и их комбинаций, продолжают быстро накапливаться. В добавление к этому появилось несколько работ, в которых даны методические указания по использованию этой шкалы (Sattler, 1988; Glutting, & Kaplan, 1990; Kamphaus, 1993). Современная редакция Стэнфорд—Бине отражает истинный прогресс в конструировании шкалы. *SB*-IV обеспечивает необходимую гибкость, позволяя пользователям оценивать отдельные способности в соответствии с конкретными целями тестирования. Наконец, эта версия шкалы гораздо лучше согласуется с современными теоретическими представлениями о природе интеллекта и свежими данными исследований в этой области (см. главу 11).

Шкалы Векслера

Разработанные Дэвидом Векслером шкалы интеллекта включают несколько последовательных редакций трех шкал: для взрослых, для детей школьного возраста и для дошкольников. Помимо их использования для измерения общего интеллекта век-

слеровские шкалы пробовали применять в качестве вспомогательного средства психиатрического диагноза. Опираясь на наблюдение, что повреждения мозга, психотические обострения и эмоциональные расстройства могут избирательно воздействовать на интеллектуальные функции, Д. Векслер и другие медицинские психологи утверждали, что сравнительный анализ выполнения пациентом разных субтестов мог бы пролить свет на специфику психического расстройства. Проблемы и результаты, относящиеся к такому анализу профиля шкал Векслера, будут рассмотрены в главе 17 как пример использования тестов в условиях клиники.

Об интересе к шкалам Векслера и широте их применения свидетельствуют несколько тысяч посвященных им публикаций, появившихся к настоящему времени. Помимо обычных обзоров по тестам в *Ежегодниках психических измерений* исследования, касающиеся шкал Векслера, периодически освещаются в журналах (Guertin, Frank, & Rabin, 1956; Guertin, Ladd, Frank, Rabin, & Hiester, 1966; Guertin, Ladd, Frank, Rabin, & Hiester, 1971; Guertin, Rabin, Frank, & Ladd, 1962; T. D. Hill, Reddon, & Jackson, 1985; Littell, 1960; Rabin, & Guertin, 1951; I. L. Zimmerman, & Woo-Sam, 1972) и обобщены в нескольких книгах (например, Forster & Matarazzo, 1990; Gyurke, 1991; Kamphaus, 1993; Kaufman, 1979, 1990, 1994; Sattler, 1988, 1992).

Прошлое и настоящее векслеровских шкал интеллекта. Первая форма шкал Векслера, известная как шкала интеллекта Векслера—Белльвью, была опубликована в 1939 г. Одной из главных целей подготовки этой шкалы была разработка теста интеллекта, пригодного для тестирования взрослых людей. Представляя впервые эту шкалу, Д. Векслер (Wechsler, 1939) отмечал, что доступные ранее тесты интеллекта разрабатывались главным образом для школьников и адаптировались для взрослых добавлением более трудных заданий того же типа. Содержание таких тестов часто не представляло никакого интереса для взрослых людей. Если задания теста не обладают хотя бы минимумом очевидной валидности, то практически невозможно установить должный раппорт со взрослыми испытуемыми. Многим заданиям теста интеллекта, специально составленным с учетом повседневных занятий ребенка школьного возраста, явно не хватает очевидной валидности с точки зрения большинства взрослых.

Ориентировка большинства тестов на скорость выполнения может также ставить в невыгодные условия пожилых людей. Кроме того, Д. Векслер считал, что в традиционных тестах интеллекта неоправданно большое значение придавалось относительно шаблонным манипуляциям словами. Он обратил внимание коллег на неприменимость норм умственного возраста к взрослым и указал на то, что прежние выборки стандартизации для индивидуальных тестов интеллекта включали лишь незначительное число взрослых.

Стремление преодолеть все эти недостатки и привело к разработке первой шкалы Векслера—Белльвью. По форме и по содержанию эта шкала служит базисной моделью для всех последующих векслеровских шкал интеллекта, каждая из которых, в свою очередь, вносила некоторые усовершенствования в предшествующую ей версию. В 1949 г. была подготовлена Векслеровская шкала интеллекта для детей (*WISC*) как расширение шкалы Векслера—Белльвью в сторону более низких возрастных уровней (Seashore, Wesman, & Doppelt, 1950). Многие задания были взяты непосредственно из теста для взрослых, и в каждый субтест были добавлены более легкие задания того же типа. В 1955 г. шкала Векслера—Белльвью была вытеснена Векслеровской шкалой интеллекта для взрослых (*WAIS*), свободной от некоторых технических не-

достатков прежней шкалы, касающихся объема и репрезентативности нормативной выборки, а также надежности субтестов. В 1967 г. семейство тестов Векслера пополнилось еще одним, «самым младшим ребенком» — Векслеровской шкалой интеллекта для дошкольников и младших школьников (*WPPSI*), первоначально задуманной для детей от 4 до 6,5 лет как расширение нижней области возрастного диапазона *WISC*, которая предназначалась для детей от 5 до 15 лет.

Разработка *WISC* с самого начала была отмечена известными противоречиями, так как Векслер приступил к созданию своих тестов отчасти из-за острой потребности в такой шкале для измерения интеллекта взрослых, которая *не* была бы простым расширением имеющихся на тот момент шкал для детей в сторону более высоких возрастных уровней. Первая редакция *WISC* была фактически полностью раскритикована за недостаточную ориентацию ее содержания на детей. В пересмотренной редакции этой шкалы (*WISC-R*), изданной в 1974 г. и предназначавшейся для детей от 6 до 16 лет, ориентированные на взрослых задания были заменены или изменены таким образом, чтобы приблизить их содержание к обычному детскому опыту. В арифметическом субтесте, например, в условиях задачи «сигары» были заменены «конфетами». Другие изменения состояли в исключении заданий, которые могли быть в разной степени знакомы отдельным группам детей, и включении большего количества женских и негритянских персонажей в наглядный материал субтестов. Ряд субтестов пришлось удлинить в целях повышения их надежности. Кроме того, были внесены некоторые усовершенствования в процедуры проведения теста и подсчета баллов.

Описание шкал. К настоящему времени каждая из трех шкал Векслера подверглась хотя бы одной, а то и нескольким переработкам. Современных версий шкал, опубликованных под именем Дэвида Векслера уже после его смерти в 1981 г., три: Пересмотренная шкала интеллекта взрослых Векслера (*WAIS-R* — Wechsler, 1981), охватывающая возрастной диапазон от 16 до 74 лет; Векслеровская шкала интеллекта для детей — Третья редакция (*WISC-III* — Wechsler, 1991), предназначенная для детей от 6 лет до 16 лет 11 месяцев; Пересмотренная Векслеровская шкала интеллекта для дошкольников и младших школьников (*WPPSI-R* — Wechsler, 1989), покрывающая теперь возрастной диапазон от 3 лет до 7 лет 3 месяцев. Третью редакцию шкалы интеллекта взрослых (*WAIS*), работа по усовершенствованию которой велась с 1992 г., предполагается подготовить к 1997 г.

WAIS-R. WISC-III и *WPPSI-R* имеют много общих черт, включая основную организацию Вербальной и Невербальной шкал, каждая из которых состоит минимум из пяти (а максимум из семи) субтестов и дает отдельные показатели в единицах стандартного *IQ.* Индивидуальные показатели по всем 10 систематически проводимым субтестам (11 для *WAIS-R*) объединяются в Полную шкалу *IQ* (*Full Scale IQ*), которая имеет то же среднее и стандартное отклонение ($M = 100$, $SD = 15$), что и две подшкалы — Вербальная и Невербальная. Из 17 различных видов субтестов, используемых в *WAIS-R, WISC-III* и *WPPSI-R*, восемь (5 вербальных и 3 невербальных) являются общими для всех трех шкал. При применении этих шкал вербальные и невербальные субтесты чередуются и предъявляются в заранее установленной последовательности, своей для каждой шкалы.

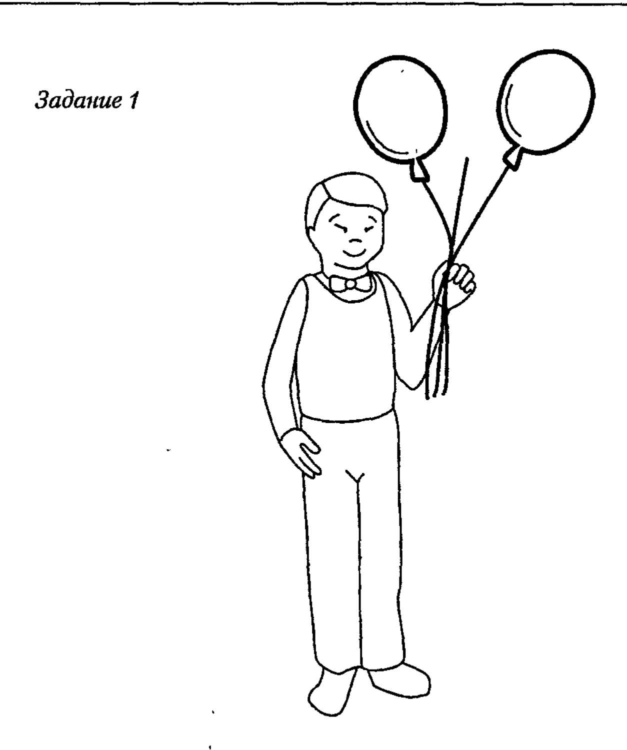
Субтест «Осведомленность» — первый вербальный субтест, предъявляемый во всех трех шкалах и служащий хорошим средством установления раппорта с тестируемым. Было затрачено немало усилий, чтобы избежать в нем вопросов, касающихся специ-

альных знаний. Его первые задания достаточно легки для того, чтобы с ними справились подавляющее большинство тестируемых, если только они не страдают умственной отсталостью или нарушением ориентации в действительности. В таких случаях тестирующий может быстро принять решение о прекращении тестирования. Вопросы субтеста «Осведомленность» в версиях *WAIS-R* и *WISC-III* касаются фактов, о которых большинство живущих в США скорее всего имело шанс узнать, например: «Какой месяц наступает перед декабрем?» или «Кем был Марк Твен?» В версии *WPPSI-R* предлагаются аналогичные вопросы, хотя и на более низком уровне трудности. На самом деле, эта версия начинается с заданий, предъявляемых в изобразительной форме, которые требуют только показать правильный ответ. Например, при предъявлении картинки с изображением нескольких бытовых предметов ребенка могут спросить, какой из них используется для уборки. Субтест «Арифметический» — еще одна вербальная мера, демонстрирующая широкий диапазон трудности на группе шкал Векслера. В самых легких арифметических заданиях *WPPSI-R* требуется показать только один предмет в ряду, иллюстрирующем количественное понятие (такое, как «самый маленький» или «больше»). Более сложные задания могут быть связаны с вычислениями или решением арифметических задач, самые трудные из которых требуют хорошего усвоения дробей.

Невербальные субтесты (или, по-другому, субтесты действия) шкал Векслера обычно требуют манипулирования различными объектами, такими как части разрезанных фигур и кубики, или визуального обследования печатных материалов наподобие картинок или набора символов. Все они устанавливают временны́е лимиты для тестируемого, которому в большинстве случаев начисляются к тому же дополнительные баллы за скорость. В противоположность этому, в Вербальной шкале только один субтест (Арифметический) является скоростным. Субтест «Недостающие детали» — невербальный субтест, используемый во всех трех шкалах Векслера; он требует от тестируемого определить, какой важной части недостает в изображениях знакомых предметов или обычных сцен. Задания для ранних возрастов рассчитаны на простое визуальное обследование, — например, как в случае предъявления изображения животного с отсутствующей конечностью. В более трудных заданиях для установления недостающего элемента необходимо дедуктивное рассуждение, специальное знание или то и другое вместе. На рис. 8–3 показаны два относительно легких задания на установление недостающих деталей, аналогичных используемым в шкалах Векслера.

Сокращенные шкалы. Со времени выхода в свет первой шкалы Векслера—Белльвью было предложено множество *сокращенных шкал (abbreviated scales) или кратких форм (short forms)* тестов Векслера. Цель этих сокращенных шкал — существенно сократить время тестирования при получении показателя *IQ* в Полной шкале, который можно оценить на основе опубликованных норм. Самый простой способ построения таких более коротких форм — опустить некоторые из субтестов и пропорционально распределить показатели. Кроме того, сокращенные шкалы создавали путем уменьшения числа заданий в субтестах.

То, что некоторые комбинации субтестов имеют корреляции с показателями *IQ* Полной шкалы, превышающие 0,90, стимулировало разработку и использование сокращенных шкал для целей быстрого отсеивания обследуемых. Были проведены обширные исследования, чтобы установить наиболее эффективные комбинации двух, трех, четырех и пяти субтестов в предсказании *IQ* по Вербальной, Невербальной и



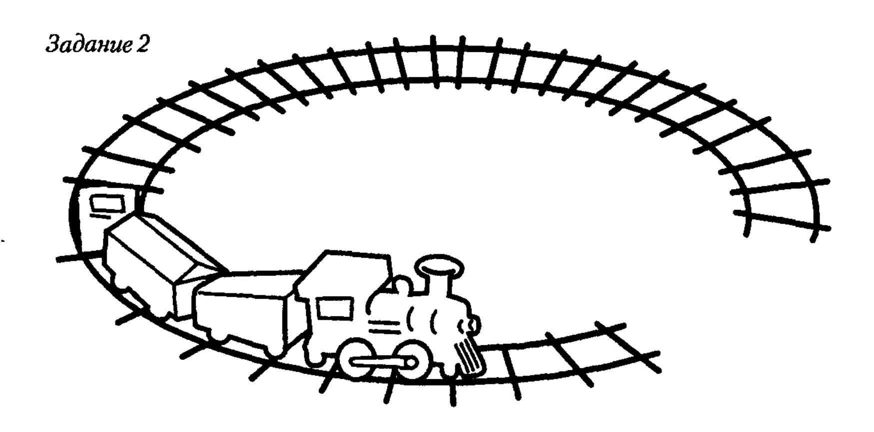


Рис. 8–3. Два задания на установление недостающих деталей, аналогичные используемым в Пересмотренной Векслеровской шкале интеллекта для дошкольников и младших школьников
(С любезного разрешения The Psychological Corporation)

Полной шкалам (Matarazzo, 1972; McCusker, 1994; Sattler, 1988, 1992). По большей части в этих работах использовали данные стандартизации, но несколько исследований было проведено на специфических популяциях, таких как пациенты психиатрических клиник и умственно отсталые.

На составление и проверку кратких форм шкал Векслера было затрачено чрезвычайно много сил и энергии. Несмотря на это, неоднократно поднимались вопросы о качестве существующих процедур, используемых при получении сокращенных шкал из полных (Silverstein, 1990). Например, допущение о том, что нормы исходной Полной шкалы применимы к пропорционально распределенным суммарным показателям по кратким шкалам, может быть не всегда обоснованным. Кроме того, многие качественные наблюдения, которые делает возможным применение индивидуальной шкалы, теряются при использовании сокращенных шкал. Поэтому, вероятно, нецелесообразно использовать такие сокращенные версии кроме тех случаев, когда требуются грубые инструменты отсеивания.

Нормы и получение показателей. К формированию выборок стандартизации самых последних шкал Векслера подходили с особой осторожностью, чтобы обеспечить их репрезентативность. Нормативные выборки включали примерно по 2000 испытуемых для каждой шкалы, уравненных по полу и распределенных по соответствующим возрастным группам. Испытуемых отбирали таким образом, чтобы выборки как можно точнее соответствовали данным последних отчетов Бюро переписи населения США, доступным на момент стандартизации, с учетом таких переменных, как географический район, расовая или этническая принадлежность, профессиональный и образовательный уровень. В случае детей учитывался профессиональный уровень родителей. С каждым последующим пересмотром шкал, переменные, используемые в плане стратифицированного отбора испытуемых, несколько изменялись в направлении, обеспечивающем бо́льшую инклюзивность выборок стандартизации. Например, категория этнической принадлежности при стандартизации *WISC-III* включала четыре группы (белые, черные, испаноязычные и прочие), тогда как стандартизация более ранней версии *WISC-R* проводилась на выборке, стратифицированной по этой переменной только на две группы (белые или небелые). К тому же в отличие от более ранних шкал, выборка стандартизации *WISC-III* включала в качестве особо выделенной репрезентативную группу учащихся, получающих специальные услуги в условиях школы, такие как обучение детей-инвалидов и работа с одаренными детьми.

Популярность шкал Векслера, которые в настоящее время являются наиболее широко используемыми индивидуальными тестами интеллекта, стала причиной ряда исследований, задуманных с целью расширения их пригодности. Так, в составе серии нормативных исследований, проводимых на пожилых американцах в клинике Мэйо (Mayo Clinic), с целью получения нормативных данных для *WAIS-R* за пределами самой старшей возрастной группы выборки стандартизации были обследованы 222 человека в возрасте от 56 до 97 лет (Ivnik et al., 1992). В другом исследовании были составлены нормы на основе данных, полученных при обследовании 130 человек в возрасте старше 75 лет (Ryan, Paolo, & Brungardt, 1990).

Первичные показатели по каждому субтесту шкал Векслера преобразуются в стандартные показатели со средним значением, равным 10, и $SD = 3$. Таким образом, все нормированные показатели субтестов выражаются в сравнимых единицах. Затем эти показатели субтестов, соответствующих Вербальной, Невербальной и Полной шка-

лам, складываются и преобразуются в стандартные показатели со средним, равным 100, и *SD* = 15, называемые «стандартным *IQ*». Кроме того, *WISC-III* дает четыре дополнительных, основанных на результатах факторного анализа, балльных индекса (*index scores*), а именно: Индекс Вербального Понимания (*Verbal Comprehension* [*VCI*]), Индекс Перцептивной Организации (*Perceptual Organization* [*POI*]), Индекс Внимательности (*Freedom from Distractibility* [*FDI*]) и Индекс Скорости Обработки Информации (*Processing Speed* [*PSI*]). Состав этих индексов имеет тесное сходство с составом факторов, типично выделяемых в результате факторного анализа более ранней версии WISC-R многими независимыми исследователями. Эти новые индексы основаны на комбинациях двух или четырех субтестов и имеют среднее, равное 100, и *SD* = 15. Каждая из трех шкал Векслера дает информацию, необходимую для оценки индивидуального результата по любым отдельным субтестам и их группам исходя из соответствующих возрастных норм.

Надежность. Векслеровские шкалы предоставляют информацию о коэффициентах надежности эквивалентных половин для показателя каждого субтеста,[1] балльного индекса и *IQ* по всем возрастным группам. Для всех шкал Векслера соответствующие коэффициенты надежности *IQ* Полной шкалы колебались от 0,90 до 0,98, *IQ* Вербальной шкалы — от 0,86 до 0,97 и *IQ* Невербальной шкалы — от 0,85 до 0,94. Четыре балльных индекса *WISC-III* получили коэффициенты надежности эквивалентных половин в диапазоне от 0,80 до 0,95. Как можно было ожидать, надежность субтестов оказалась несколько ниже. Что касается трех обсуждаемых нами шкал Векслера, надежность входящих в них субтестов колебалась от 0,52 до 0,96, при этом подавляющее большинство субтестов дало коэффициенты надежности выше 0,70. Надежность субтестов особенно важно учитывать при оценивании значимости различий между показателями субтестов одного и того же человека, как в случае анализа профиля (J. H. Kramer, 1990, 1993; Sattler, 1988, 1992). В руководствах к шкалам Векслера также приводятся стандартные ошибки измерения для всех видов показателей. Для *IQ* Вербальной шкалы такие ошибки варьируют от 2,50 до 4,98 единиц, для *IQ* Невербальной шкалы — от 3,67 до 4,97 единиц, а для *IQ* Полной шкалы все они меньше 4,00 единиц. Таким образом, мы можем, к примеру, заключить: шансы того, что истинное значение *IQ* Полной шкалы у конкретного человека отстоит не больше чем на 4 единицы от *IQ*, полученного им по Полной шкале, составляют примерно 2 : 1.

Данные по ретестовой надежности (устойчивости) показателей шкалы Векслера собирались более тщательно, при каждом ее пересмотре. Коэффициенты устойчивости, в тенденции, выше для взрослых, чем для детей. Ретестовые исследования неизменно показывают прирост от 2 до 13 единиц в различных показателях *IQ* от первого ко второму тестированию, интервал между которыми составляет от 12 дней до 9 недель; при этом *IQ* Полной шкалы типично возрастает на 5–7 единиц. Такой ожидаемый эффект упражнения, хотя и незначительный, следует принимать в расчет при повторном тестировании испытуемых через короткий промежуток времени.

Руководства по *WPPSI-R* и *WISC-III* — самые последние в серии руководств, последовательно совершенствуемых со временем. Среди многих заслуживающих внима-

---

[1] За исключением тех субтестов, для которых коэффициент надежности эквивалентных половин неприменим, т. е. субтестов «Цифровые символы» (*Digit Symbol*), «Кодирование» (*Coding*), «Дом животного» (*Animal Pegs*), «Поиск символов» (*Symbol Search*) и «Повторение цифр» (*Digit Span*).

ния особенностей этих руководств — включение коэффициентов надежности оценщика для субтестов, которые требуют при начислении баллов опоры на субъективные суждения. Эти данные свидетельствуют о том, оценивание ответов по таким субтестам могут производить с приемлемой надежностью только получившие специальную подготовку и практический опыт работы пользователи теста. Еще одно новшество в руководствах по этим шкалам — предоставление довольно большого количества данных, касающихся внутрииндивидных различий показателей. В добавление к таблицам, показывающим степень различий, необходимых для достижения статистической значимости, в этих руководствах приводятся частоты различий, обнаруженных внутри выборки стандартизации. Информация такого рода имеет особую ценность при клиническом использовании рассматриваемых шкал (см. главу 17).

Валидность. Нигде улучшение в ходе последовательного совершенствования руководств к шкалам Векслера не было столь выраженным, как в области валидности. В 1981 г., когда была опубликована *WAIS-R*, в руководстве к этой шкале не было никаких данных о ее валидности кроме результатов двух корреляционных исследований, в которых оценивались связи между показателями *WAIS-R* и более ранних шкал Векслера. Сведений о валидности в нем было даже меньше, чем в руководствах по *WPPSI* и *WISC-R*, которые по крайней мере содержали данные о корреляции показателей этих шкал с показателями других тестов интеллекта, таких как шкала Стэнфорд—Бине. Однако ограниченный охват данных о валидности в старых руководствах к шкалам Векслера в какой-то мере компенсировался значительным объемом опубликованных независимых исследований валидности всех этих шкал.[1] Отчасти недостаточное внимание к сведениям о валидности в руководствах к шкалам Векслера было вызвано убеждением Дэвида Векслера в том, что задачи в его шкалах охватывают диапазон специфических способностей, вполне достаточный для обеспечения валидной оценки общего интеллекта (Wechsler, 1958; Zachary, 1990).

Трактовка валидности Д. Векслером отражала, по существу, ориентацию на описание содержания, хотя и содержала некий подтекст, связанный с подходом к валидизации через идентификацию конструкта с минимальным привлечением подтверждающих данных. Например, корреляции между шкалами Векслера и другими глобальными мерами интеллекта, такими как шкала Стэнфорд—Бине, группируются около 0,80. В дополнение к этому, результаты факторного анализа полученных с помощью шкал Векслера данных, проводимого независимыми исследователями на протяжении целого ряда лет, обнаружили удивительную согласованность. Во всех группах почти неизменно выделялся вербальный фактор и фактор перцептивной организации (или невербальный). В шкалах для более старших детей и взрослых типично выявлялись дополнительные факторы памяти и/или внимания. Использование шкал Векслера в профессиональном и образовательном отборе было оправдано, в известной степени, данными о различиях в ожидаемом направлении между разными группами.

Напротив, данные о всех типах валидности самых последних шкал Векслера представлены в изобилии. В руководствах по *WPPSI-R* и *WISC-III* обсуждению валидности посвящено 21 и 38 страниц соответственно, в противоположность 2 страницам,

[1] См. Dean, 1977, 1979, 1980; Gutkin, & Reynolds, 1981; G. P. Hollenbeck, & Kaufman, 1973; Karnes, & Brown, 1980; Kaufman, 1975; Kaufman, & Hollenbeck, 1974; Leckliter, Matarazzo, & Silverstein, 1986; Silverstein, 1982a, 1982b; Waller, & Waldman, 1990.

отведенным этой теме в руководстве по *WAIS-R*. Данные, относящиеся к валидизации конструкта, получены путем интеркорреляций субтестов и факторного анализа показателей. Усредненные (по всем возрастным группам) интеркорреляции между Вербальной и Невербальной шкалами, полученные на выборке стандартизации, составляют 0,74 для *WAIS-R*, 0,66 для *WISC-III* и 0,59 для *WPPSI-R*; величина коэффициентов свидетельствует о наличии значительного общего фактора, что подтверждается большинством результатов факторного анализа трех этих шкал.

Исследования с применением факторного анализа девяти возрастных групп выборки стандартизации *WAIS-R* по большей части говорят о том, что наилучшим объяснением корреляций между 11 субтестами служит трехфакторная модель. Выделяемые факторы, которые, по-видимому, можно распространить на различные типы выборок, включают Вербальное понимание, Перцептивную организацию и Память/Внимательность (Leckliter et al., 1986; Waller, & Waldman, 1990). Анализ того же типа, проведенный с данными стандартизации *WPPSI-R* и описанный в руководстве и в других источниках, дает двухфакторное решение, согласующееся с организацией субтестов в Вербальную и Невербальную шкалы (Blaha, & Wallbrown, 1991; LoBello, & Gulgoz, 1991; B. J. Stone, Gridley, & Gyurke, 1991). С другой стороны, данные по *WISC-III*, с самого начала подвергавшиеся как разведочному, так и подтверждающему факторному анализу, результаты которого описаны в руководстве к этой шкале, лучше всего согласуются с четырехфакторной моделью, включающей такие факторы, как Вербальное понимание, Перцептивная организация, Внимательность и Скорость обработки информации. Эти четыре фактора и были введены в состав стандартных показателей *WISC-III*.

В руководствах по *WPPSI-R* и *WISC-III* также представлены данные о валидности из многочисленных исследований (хотя и с мало подходящими выборками), в которых устанавливаются корреляции этих двух шкал с другими индивидуально проводимыми тестами. В случае *WISC-III* приводятся еще корреляции с групповыми тестами достижений и школьными оценками. Кроме того, диагностическая или прогностическая (в отношении критерия) полезность *WISC-III* и *WPPSI-R* изучалась в серии исследований особых групп, включая одаренных, умственно отсталых, испытывающих трудности в обучении и другие типы детей.

Заключительные замечания по шкалам Векслера

Последовательные редакции трех шкал Векслера отражают возрастающий уровень изощренности и опыта в конструировании тестов, соответствующий сменявшимся десятилетиям, в которые они разрабатывались. По сравнению с другими индивидуально проводимыми тестами главные достоинства этих шкал связаны с объемом и репрезентативностью выборок стандартизации, особенно для совокупностей взрослых и детей дошкольного возраста, а также с техническими характеристиками процедур, применяемых при их конструировании. Следует особо отметить уровень рассмотрения вопросов надежности и валидности в руководстве по *WISC-III*. Популярность шкал Векслера гарантирует им постоянно расширяющуюся базу исследований, по крайней мере на какое-то время. К тому же для их пользователей доступно множество вспомогательных материалов, таких как программы машинной интерпретации данных, руководства для подготовки специалистов по тестированию (например, Fantuzzo, Blakey, & Gorsuch, 1989) и руководства по интерпретации результатов тести-

рования (например, Kaufman, 1994; Nicholson, & Alcorn, 1994; Whitworth, & Sutton, 1993). Однако некоторые критики отмечают, что даже самые последние, наиболее усовершенствованные версии шкал Векслера вскоре могут устареть и стать ненужными в свете современных требований к связям между инструментами оценивания и стратегиями вмешательства (Shaw, Swerdlik, & Laurent, 1993; Sternberg, 1993). В этом отношении самой уязвимой стороной всех шкал Векслера была и остается слабость их теоретического обоснования, препятствующая нахождению прочной и связной основы для интерпретации результатов тестирования. Кроме того, структура этих шкал, по-видимому, основана на предположении, что области способности, «простукиваемые» входящими в них субтестами, остаются одними и теми же (судя по внешнему сходству тестовых материалов и задач) на всех возрастных уровнях. Тем не менее это предположение может оказаться несостоятельным в свете того, что нам уже известно о возрастных изменениях интеллекта на протяжении жизни (см. главу 11).

Шкалы Кауфмана

Шкалы Кауфмана — это клинические инструменты индивидуального применения, предназначенные для использования во многих областях, для которых были разработаны и в которых традиционно применялись такие тесты, как шкалы Стэнфорд—Бине и Векслера (Kaufman, & Kaufman, 1983a, 1983b, 1990, 1993). Разработанные в период с 1980-х по начало 1990-х гг. шкалы Кауфмана вобрали в себя последние достижения в области конструирования тестов. Оценочная батарея Кауфмана для детей (*Kaufman Assessment Battery for Children* [*K-ABC*] — Kaufman, & Kaufman, 1983a, 1983b) и особенно Тест интеллекта подростков и взрослых Кауфмана (*Kaufman Adolescent and Adult Intelligence Test* [*KAIT*] — Kaufman, & Kaufman, 1993) представляют собой попытки со стороны их авторов, — участвовавших, кстати, в разработке *WISC-R*, — преодолеть чисто эмпирическую позицию, преобладавшую при создании более ранних шкал интеллекта. Они стремились создать инструменты, которые по замыслу были бы привязаны к развивающимся теориям интеллекта, включали соответствующие возрастному развитию задачи и давали полезную информацию для разнообразных ситуаций оценивания.

Оценочная батарея Кауфмана для детей (*K-ABC*)

Сущность и построение. Конструирование *K-ABC* началось с определения подлежащих оцениванию конструктов. В соответствии с генеральной линией когнитивной психологии главное внимание было уделено обработке информации. Выбранный в данном случае подход разграничивает параллельную обработку информации, оцениваемую семью субтестами, и последовательную обработку, оцениваемую тремя субтестами (J. P. Das, 1984; Das, Kirby, & Jarman, 1975, 1979; Das, & Molloy, 1975; Kaufman, & Kaufman, 1983b, chap. 2; Luria, 1966). Субтесты шкалы «Параллельная обработка информации» требуют синтеза и организации пространственных образов и зрительно воспринимаемого содержания, которые могут обозреваться как нечто целое. Субтесты шкалы «Последовательная обработка информации» требуют сериальной или временной организации; они предполагают использование вербального, числового и зрительно воспринимаемого содержания, а также кратковременной памяти. Несколько

задач, представленных в объединенной шкале «Умственная обработка информации», имеют сходство с задачами, используемыми в нейропсихологическом обследовании (см. главу 17), и были выбраны как раз по этой причине.

Эта батарея включает, кроме того, «Шкалу достижения», содержащую шесть субтестов. Несмотря на то, что входящие в эту шкалу субтесты оценивают умения читать и выполнять арифметические действия, знание слов и общую осведомленность, их конструировали *вовсе не* для измерения фактуальных знаний, которым учат в школе. Они гораздо более похожи на задачи, включаемые в традиционные тесты интеллекта или способностей, чем на задания традиционных тестов учебных достижений. В арифметическом тесте, например, ребенок рассматривает серию картинок о семье, пришедшей в зоопарк, и должен реагировать считая на каждой картинке изображенные объекты или выполняя с ними простые числовые операции. Понимание прочитанного демонстрируется выполнением действий, описанных в каждом предложении, которое ребенок читает.

K-ABC была стандартизована на национальной выборке, включавшей 2000 детей в возрасте от 2,5 до 12,5 лет. В дополнение к этому было протестировано несколько групп черных и белых детей с целью разработки социокультурных норм с учетом расы и образования родителей — полезного дополнения для более адекватной интерпретации результатов. К тому же *K-ABC* изначально создавалась таким образом, чтобы ее можно было приспособить к потребностям тестирования особых групп, таких как дети-инвалиды и дети, принадлежащие к культурным и языковым меньшинствам, а также использовать как вспомогательное средство при диагностике трудностей в обучении (Kamphaus, Kaufman, & Harrison, 1990). Эта батарея дает четыре общих показателя: «Последовательная обработка информации» (*Sequential Processing*), «Параллельная обработка информации» (*Simultaneous Processing*), «Умственная обработка информации» (*Mental Processing Composite*) — совокупный показатель, объединяющий первые два, и «Достижение» (*Achievement*). Каждый из них представляет собой стандартный показатель со средним, равным 100, и *SD* = 15.

Общая оценка. *K-ABC* обладает многими достоинствами как технического, так и практического характера.[1] В соответствии с духом времени, распространенной тенденции относить детей к той или иной категории на основе единственной числовой оценки, такой как *IQ*, здесь ставится надежный заслон благодаря использованию множественных показателей, разных вариантов анализа профиля и диагностических интерпретаций, особенно удачно описанных в главах 5 и 6 *Руководства по интерпретации результатов K-ABC* (*Interpretive Manual* — Kaufman, & Kaufman, 1983 b). В главе 6 этого руководства дана блестящая иллюстрация цикла порождения и проверки гипотезы, который составляет сущность клинического подхода к диагностике. Кроме того, пытаясь рассеять некоторые неверные представления, получившие широкое распространение, создатели этой батареи открыто заявляют во вступительной главе *Руководства по интерпретации...* (Kaufman, & Kaufman, 1983b, p. 20–24), что *K-ABC не* является «мерилом врожденных или неизменных способностей», добавляя при этом, что «все когнитивные задачи рассматриваются в качестве критериев того, чему индиви-

[1] Что касается критических обзоров и дискуссий, см. прежде всего T. L. Miller (1984). См. также Anastasi (1984a, 1985c), Coffman (1985), Kamphaus (1990), Kline, Snyder, & Castellanos (1996), Page (1985).

дуум научился». Они откровенно предупреждают, что *K-ABC*, подобно любому другому тесту, нельзя считать «завершенной тестовой батареей» и следует дополнять другими инструментами в соответствии с индивидуальными потребностями.

Но вопреки предостерегающим заявлениям авторов, употребление ими термина «тесты достижений», возможно, было неудачным выбором из-за преобладания ошибочных представлений об отношении между тестами способностей и тестами достижений. Тест можно уверенно отнести к категории тестов достижений, когда он тесно связан со специфическим, поддающимся четкому определению, содержанием обучения, которое тестируемые, предположительно, должны пройти. Однако этого нельзя сказать в отношении тестов, обозначенных как «тесты достижения» в батарее K-ABC, при создании которой прилагались специальные усилия, чтобы отделить ее тесты от специфических знаний, приобретаемых в классе. Фактически, внутри континуума развиваемых способностей, эти тесты гораздо ближе к концу способностей (*aptitude*), чем к концу достижений (*achievement*), — вывод, подтверждаемый интеркорреляциями субтестов. Поэтому вряд ли можно считать оправданным употребление терминологии, которая приобрела дополнительные значения, несет в себе непреднамеренные импликации и поддерживает распространенные заблуждения.

Формулировка в явном виде теоретической основы как руководства для составления спецификации задач и разработки заданий в *K-ABC* явилась желанным нововведением, согласующимся с принципами конструирования хороших тестов. И хотя прошло уже более десяти лет с момента выпуска этой батареи, остаются вопросы по поводу того, была ли выбранная ее создателями теоретическая ориентация наилучшей для достижения намеченных целей. В частности, высказывались сомнения в том, что различение параллельной и последовательной обработки информации может служить основой для понимания результатов выполнения *K-ABC*, и приводились доводы в пользу того, что два набора субтестов, названных в соответствии с таким различением, вполне можно было бы охарактеризовать как тесты вербального и невербального рассуждения (J. P. Das, 1984; Goetz, & Hall, 1984; A. R. Jensen, 1984; Keith, 1985; Keith, & Dunbar, 1984).

С другой стороны, уже накопленные данные исследований по *K-ABC* говорят о сходстве ее общих показателей с показателями *WISC-R* в том, что касается их прогностической валидности и того, в какой степени они измеряют «общий интеллект» (Kamphaus, 1990). Вследствие меньшей зависимости от вербальных навыков, *K-ABC* может быть предпочтительной мерой для детей с ограниченным знанием английского языка или с нарушениями слуха. Сбалансированное изложение достоинств и ограничений этого относительно нового инструмента можно найти в работе *Clinical and Research Applications of the K-ABC* (Kamphaus, & Reynolds, 1987, chap. 8).

Тест интеллекта подростков и взрослых Кауфмана (KAIT)

Сущность и построение. *KAIT* (Kaufman, & Kaufman, 1993) разрабатывался как средство измерения интеллекта в возрастном диапазоне от 11 до 85 лет (или даже старше). При его создании была сделана попытка интегрировать теорию текучего и кристаллизованного интеллекта, сформулированную Хорном и Кэттеллом (Horn, & Cattell, 1966), с представлениями других теоретиков об интеллекте взрослых людей (Golden, 1981; Luria, 1980; Piaget, 1972).

Данная батарея составлена из двух шкал. Шкала «Кристаллизованный интеллект» (*Crystallized Scale*) измеряет представления и понятия, приобретенные в процессе школьного обучения и аккультурации, тогда как шкала «Текучий интеллект» оценивает способность решать новые задачи. В состав «Основной батареи» (*Core Battery*) входит по три субтеста из каждой шкалы. Кроме того, может использоваться «Расширенная батарея» (*Expanded Battery*), предназначенная для обследования пациентов с подозрением на локальные поражения мозга, которая образуется добавлением любого из четырех специализированных субтестов. Наконец, *KAIT* включает краткий тест Психического статуса (*Mental Status*) для оценки внимания и ориентации в обстановке у тех, кто в когнитивном отношении слишком слаб, чтобы пройти обследование с помощью полной батареи.

Общая оценка. С точки зрения своих технических характеристик *KAIT*, по-видимому, в той же степени отвечает психометрическим стандартам, как и любая другая из основных интеллектуальных шкал современного поколения. Его нормативная выборка вполне адекватна, а приводимые в руководстве данные о надежности и валидности выглядят многообещающе. *KAIT* отличается относительно легкой процедурой проведения. Кроме того, руководство к нему содержит крайне полезную информацию, касающуюся осложнений при проведении и подсчете показателей (например, что делать, когда тестируемый отвечает не на английском языке).

Однако, что действительно отличает *KAIT* от других шкал интеллекта взрослых, — это та тщательность, с какой разрабатывались и проверялись более 2500 заданий, входивших в исходную совокупность. Эти задания должны были быть привлекательными для взрослых испытуемых. Предполагалось, что для их выполнения потребуются процессы решения задач, типичные для мышления на уровне формальных операций (по Пиаже), и оценочные функции планирования, которые, согласно Luria (1980) и Golden (1981), характеризуют мышление взрослых. В результате, большинство отобранных заданий оказались довольно необычными и интересными. Многие субтесты носят занимательный характер, что отражено даже в их названиях, — например, «Знаменитые лица» (*Famous Faces*), «Тайные коды» (*Mystery Codes*) и «Двусмысленности» (*Double Meanings*). Другие субтесты отличаются новыми задачами, например, субтест «Обучение ребусам» (*Rebus Learning*). В этом субтесте тестируемые выучивают слова, связанные с конкретным ребусом (рисунком), и затем «читают» устойчивые словосочетания или предложения, составленные из таких ребусов (см. пример на рис. 8–4). Решающее испытание для *KAIT*, как и для любого нового инструмента, заключается в том, будет ли его привлекательность достаточной, чтобы вызвать к нему интерес исследователей и практиков, работа которых только и может привести к созданию богатой базы данных длительного пользования.

Краткий тест интеллекта Кауфмана (*K-BIT*)

Краткий тест интеллекта Кауфмана (*K-BIT* — Kaufman, & Kaufman, 1990) создавался как быстрый инструмент отсеивания, оценивающий уровень интеллектуальной деятельности. Данный тест, хотя и относится к категории индивидуальных, настолько прост, что его может проводить специалист среднего звена (*technician*). *K-BIT* охватывает возрастной диапазон от 4 до 90 лет. Нормирован одновременно с *KAIT* примерно на 20 % выборки стандартизации последнего, состоявшей из 2000 испытуемых.

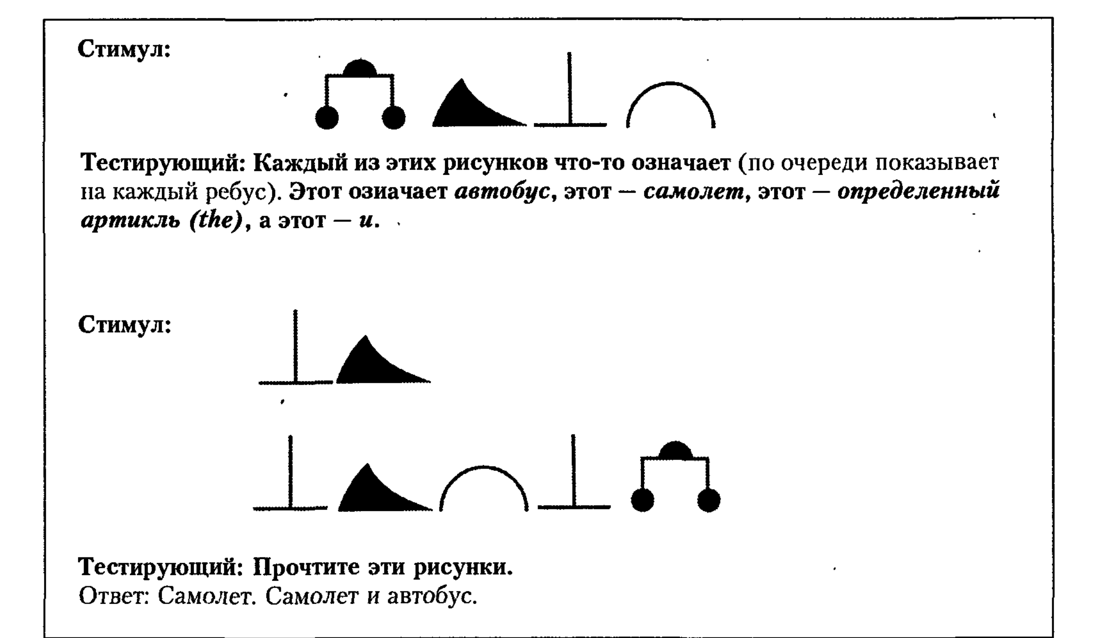


Рис. 8–4. Пример субтеста «Обучение ребусам» из Шкалы интеллекта подростков и взрослых Кауфмана

(Из Kaufman & Kaufman, 1993, p. 5. Copyright © 1993 by American Guidance Service, Inc. Воспроизведено с разрешения издателя)

K-BIT не является сокращенной версией одной из шкал Кауфмана (*K-ABC* или *KAIT*). Он состоит из одного вербального субтеста, включающего 45 заданий «Экспрессивного словаря» (*Expressive Vocabulary*) и 37 «Определений» (*Definitions*) и одного невербального субтеста из 48 «Матриц» (*Matrices*). Три показателя (вербальный, невербальный и составной), которые дает *K-BIT*, выражаются в единицах стандартного *IQ*, как и показатели других шкал Кауфмана. Длина субтестов *K-BIT* имеет следствием более высокие коэффициенты надежности по сравнению с коэффициентами надежности, характерными для кратких форм других шкал. Однако, что касается важного вопроса корреляции его показателей с показателями полных шкал, *K-BIT* вряд ли можно считать более совершенным, чем краткие формы других тестов интеллекта.

Дифференциальные шкалы способностей

Дифференциальные шкалы способностей (*Differential Ability Scales* [*DAS*] — C. D. Elliott, 1990a, 1990b) представляют собой пересмотренную и расширенную версию Британских шкал способностей (*British Ability Scales* [*BAS*]), разработанную в Великобритании в 1970-х гг. (Elliott, Murray, & Pearson, 1979). С современными версиями шкал Стэнфорд—Бине и Векслера *DAS* роднят общие цели классификации людей по общему уровню способностей и получение индивидуальных профилей сильных и слабых сторон их интеллектуальной деятельности. Однако в том, что касается проце-

дур и технических характеристик, *DAS* нетрадиционны, поскольку в них реализованы многие достижения психометрической теории и практики, не коснувшиеся других шкал. В этой связи заслуживает внимания утверждение автора в предисловии к руководству по *DAS*, что термины «интеллект» и «*IQ*» не входят в состав терминологии Дифференциальных шкал способностей (Elliott, 1990a, p. vi). В значительной степени структура шкалы, подсчет баллов и интерпретация результатов ориентированы на точно определяемые виды поведения (*behaviors*), которые фактически и оцениваются. Такое открытое заявление, впервые появляющееся в руководстве к шкале общих способностей, должно помочь рассеять стереотипы и ошибочные представления, связанные с широким употреблением этих терминов.

Описание. Батарея *DAS* создавалась, главным образом, для измерения специфических способностей (*specific abilities*) с приемлемой надежностью, чтобы оказывать помощь в достижении более сложных целей индивидуального оценивания, а именно дифференциальной диагностики и планирования вмешательства. Выбор задач, включенных в эту батарею, осуществлялся как по теоретическим соображениям, так и на эмпирической основе. Теоретическое обоснование *DAS* носит эклектический и гибкий характер. В основу батареи положен иерархический подход к умственным способностям, позволяющий выбирать различные уровни обобщенности и обеспечивающий широкую информационную базу для выведения гипотез об отдельных испытуемых. Эта структура прекрасно согласуется с эмпирическими данными о развитии когнитивных способностей. В отличие от более ранних шкал, втискивавших данные в теоретическую модель независимо от степени соответствия, батарея *DAS* сохранила только те составные части, для которых имеет место сходимость теоретического и эмпирического обоснований.

Как показано на рис. 8–5, DAS состоят из 20 субтестов, организованных в три главные компоненты: 1) основные субтесты, 2) диагностические субтесты и 3) тесты достижений. Названия этих тестов и субтестов описывают содержащиеся в них задачи и, в основном, не требуют пояснений. Двенадцать основных и пять диагностических субтестов составляют *когнитивную батарею (cognitive battery)*, подразделяемую внутри себя на два уровня: дошкольный и школьного возраста. На дошкольном уровне в нее входят четыре основных субтеста для детей в возрасте от 2;6 до 3;5 и шесть основных субтестов для детей в возрасте от 3;6 до 5;11.[1] На уровне школьного возраста (от 6;0 до 17;11) в батарею входят шесть основных субтестов. Для каждого возрастного уровня, путем суммирования показателей основных субтестов, находят показатель *Общей Концептуальной Способности (General Conceptual Ability* или, сокращенно, *GCA*), играющий роль общего суммарного показателя в этой батарее. Тесты, входящие в группу под названием «основные субтесты» (*core subtests*) батареи DAS, имеют высокие нагрузки по общему фактору (*g*) батареи. С другой стороны, диагностические субтесты имеют низкие корреляции с фактором *g* и не объединяются в групповые факторы; это означает, что они измеряют относительно независимые способности. Когда целесообразно использовать диагностические тесты, возможность их применения в возрасте от 2 до 5 лет зависит от возрастного уровня. Наконец, батарея

[1] При указании возраста для отделения количества лет от количества месяцев часто используют точку с запятой: «2;6» означает «2 года 6 месяцев». В этой книге такая форма записи будет использована всякий раз, когда возникнет потребность в перечислении последовательности возрастов.

	Дошкольный уровень 2;6–2;11	3;0–3;5	3;6–3;11	4;0–4;5	4;6–4;11	5	Уровень школьного возраста 6	7	8	9	→	17
Основные субтесты												
Складывание кубиков (*Block Building*)												
Вербальное понимание (*Verbal Comprehension*)							L					
Сходства картинок (*Picture Similarities*)	GGA							L				
Называние (*Naming Vocabulary*)									L			
Ранние понятия числа (*Early Number Concepts*)	H			GGA				L				
Копирование (*Copying*)												
Составление фигур (*Pattern Construction*)		H										
Воспроизведение образцов по памяти (*Recall of Designs*)												
Определения слов (*Word Definitions*)						H						
Матрицы (*Matrices*)										GGA		
Аналогии (*Similarities*)												
Последовательные и количественные рассуждения (*Sequential & Quantitative Reasoning*)						H						
Диагностические субтесты												
Подбор буквоподобных форм (*Matching Letter-Like Forms*)				H				L				
Воспроизведение цифр по памяти (*Recall of Digits*)	H											
Воспроизведение предметов по памяти (*Recall of Objects*)												
Узнавание картинок (*Recognition of Pictures*)	H									L		
Скорость обработки информации (*Speed of Information Processing*)						H						
Тесты достижений												
Основные навыки оперирования числами (*Basic Number Skills*)												
Правописание [произнесение слова по буквам] (*Spelling*)												
Чтение слов (*Word Reading*)						H						
		2;6–2;11	3;0–3;5	3;6–3;11	4;0–4;11	5	6	7	8	9	→	17

GCA = Общая Концептуальная Способность

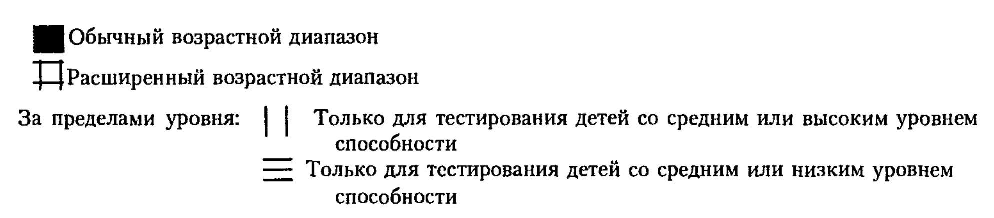


Рис. 8–5. Организация Дифференциальных шкал способностей
(С упрощениями из Elliott, 1990b, p. 4. Copyright © 1990 by The Psychological Corporation. Воспроизведено с разрешения издателя)

содержит три теста достижений, которые обычно проводятся начиная с шестилетнего возраста.

Как можно увидеть на рис. 8–5, некоторые субтесты в каждой их трех составных частей *DAS* могут предъявляться — и должным образом интерпретироваться — за пределами возрастного уровня, на который они обычно рассчитаны. Результаты выполнения субтестов, предназначенных для «расширенного возрастного диапазона» (*extended age range*) и использования «за пределами уровня» (*out of level*), могут сравниваться с нормами, основанными на выборках соответствующего возраста, собранными в процессе стандартизации *DAS*. Субтесты, входящие в категорию *«для расширенного возрастного диапазона»*, могут использоваться как дополнительные диагностические меры, когда их содержание релевантно цели обследования данного индивидуума. Например, субтест «Складывание кубиков» (*Block Building*) можно давать детям в возрасте от 3;6 до 4;11 с целью получения более полной информации о перцептивных и тонких моторных навыках, чем та, которую позволяет получить основная батарея в этом возрастном диапазоне. С другой стороны, субтесты, нормированные для тестирования *«за пределами уровня»* (на рис. 8–5 помечены буквами *H* или *L*), предназначены только для обследуемых с уровнями способности «от среднего до высокого» или «от среднего до низкого». Преимущество этой конструктивной особенности *DAS* заключается в том, что батарея позволяет проводящему тестирование с беспрецедентной точностью оценивать способности тех, кто действует на необычайно высоком или низком для своего возраста уровне.

Шкалирование и нормирование. Главная причина концептуальных и технических достоинств *DAS* состоит в том, что эта батарея вобрала в себя совокупные результаты исследовательской и теоретической работы, проделанной во время разработки ее предшественника, батареи *BAS*. Проектирование, составление и стандартизация *BAS* велись на протяжении примерно двух десятилетий, отмеченных важными достижениями в психометрической теории и практике. Таким образом *DAS* является новым инструментом, многие характеристики которого отвечают самым современным требованиям, хотя он и отражает знания и опыт, приобретенные в период с1960-х по 1980-е гг.

Стандартизацию *DAS* можно считать образцовой с точки зрения как объема выборки, так и тщательности ее комплектования. Выборка включала 3475 испытуемых, т. е. гораздо больше, чем это обычно бывает в случае стандартизации индивидуально проводимого теста. Предполагалось обеспечить ее репрезентативность относительно

изучаемой совокупности всех владеющих английским языком лиц в возрасте от 2;6 до 17;11, проживавших в США в период сбора данных (1987–1989) в домашних условиях (*noninstitutionalized*). Стратификация выборки проводилась, главным образом, по возрасту, полу, расе/этнической принадлежности, образованию родителей и географическому району проживания. Цифры, характеризующие изучаемую совокупность, основывались не на одной, отдельно взятой демографической переменной, как это бывает в типичном случае, а на составных переменных. Например, выборочное распределение белых семей с северо-востока США по образованию родителей приближалось к соответствующему распределению таких семей в совокупности населения северо-восточных штатов. Контрольные цифры рассчитывали по «сырым» данным, полученным от Бюро переписи населения США за самый последний период на момент проведения стандартизации *DAS*. Хотя выборка стандартизации и была репрезентативной относительно расового и этнического состава изучаемой совокупности (использовалось четыре категории: черные, испаноязычные, белые и прочие), дополнительно было собрано примерно 300 и 600 протоколов тестирования черных и испаноязычных детей исключительно для анализа систематической ошибки, обусловленной культурными факторами. Ученики из специальных классов, таких как классы для детей со слабыми дефектами или для особо одаренных детей, не исключались из нормативной группы, которая по замыслу исследователей должна включать полную совокупность школьников, а не только «нормальную» группу.

В *DAS* использована однопараметрическая модель теории «задание—ответ» (*IRT*),[1] что делает возможной градуировку каждого задания по уровню трудности. В результате можно использовать стратегию адаптивного тестирования, т. е. обследовать испытуемых с помощью заданий, наиболее подходящих для их уровня способности. Индивидуальный показатель основан на учете количества и уровня трудности выполненных испытуемым заданий. Эти данные наносили на общую, ненормативную шкалу, которую использовали для преобразования первичных показателей по каждому субтесту в показатели способности. Для выявления и исключения заданий, противоречащих данной модели, применялся статистический критерий согласия, основанный на соответствии между предсказанными и наблюдаемыми ответами на задания. Все это позволило создать более однородные наборы заданий.

При применении *DAS* в реальной работе стратегия адаптивного тестирования реализуется при помощи выделенных *начальных точек* (основанных на возрасте), *точек принятия решения* (основанных на результатах выполнения заданий от начальной точки до точки принятия решения) и *альтернативных правил остановки* (для каждого субтеста своих). Наборы заданий, заключенных между этими точками, определялись эмпирически, путем достижения наиболее выгодного баланса между надежностью и длиной теста. Главное достоинство стратегии адаптивного тестирования с помощью *DAS* заключено в гибкости, позволяющей тестирующему подбирать задания субтестов, подходящие для каждого тестируемого. А то, что при этом можно получить оценки способности исходя из общей шкалы трудности заданий, даже когда проводились субтесты с различными заданиями, дает пользователям дополнительное преимущество в виде допустимости сравнений показателей измеряемой данным субтестом способности у разных лиц или у одного и того человека при разных обстоятельствах. Эта характерная особенность делает *DAS*, как и другие инструменты, построенные ана-

[1] Пояснение см. в главе 7.

логичным образом, особенно подходящими для генетических исследований, использующих лонгитюдные стратегии или метод поперечных срезов.

После того как получены показатели способностей, измеряемых субтестами когнитивной батареи, их можно преобразовать в нормализованные стандартные показатели со средним 50 и *SD* = 10 (Т-показатели) или в процентильные эквиваленты. Оба типа показателей доступны для каждой возрастной группы. В тестах достижений вместо *Т*-показателей используют стандартные показатели со средним 100 и *SD* = 15, а вместо процентилей распределения по возрасту — процентили распределения по школьным классам. Для всех субтестов *DAS* можно также получить показатели в форме эквивалентных возрастов, а для тестов достижений — в форме эквивалентных классов. Эти эквиваленты указывают возраст (или класс), в котором показатель способности тестируемого соответствует медианному показателю. Поскольку используемые в *DAS* меры когнитивных способностей и меры достижений разрабатывали и нормировали одновременно, нормативные сравнения, возможные благодаря всем этим преобразованиям показателей, позволяют пользователям обращаться к широкому множеству вопросов, уместных при скрупулезном исследовании индивидуальных проблем.

Показатели основных субтестов *DAS* складываются для получения соответствующего комбинированного показателя (или показателей) на любом из возрастных уровней. Все комбинированные показатели выражаются в виде стандартных показателей со средним, равным 100, и *SD* = 15. Как показано на рис. 8–6, для самых маленьких детей (от 2;6 до 3;5) можно получить только один комбинированный показатель — показатель *GCA*; в возрастном диапазоне от 3;6 до 5;11 батарея *DAS*, в дополнение

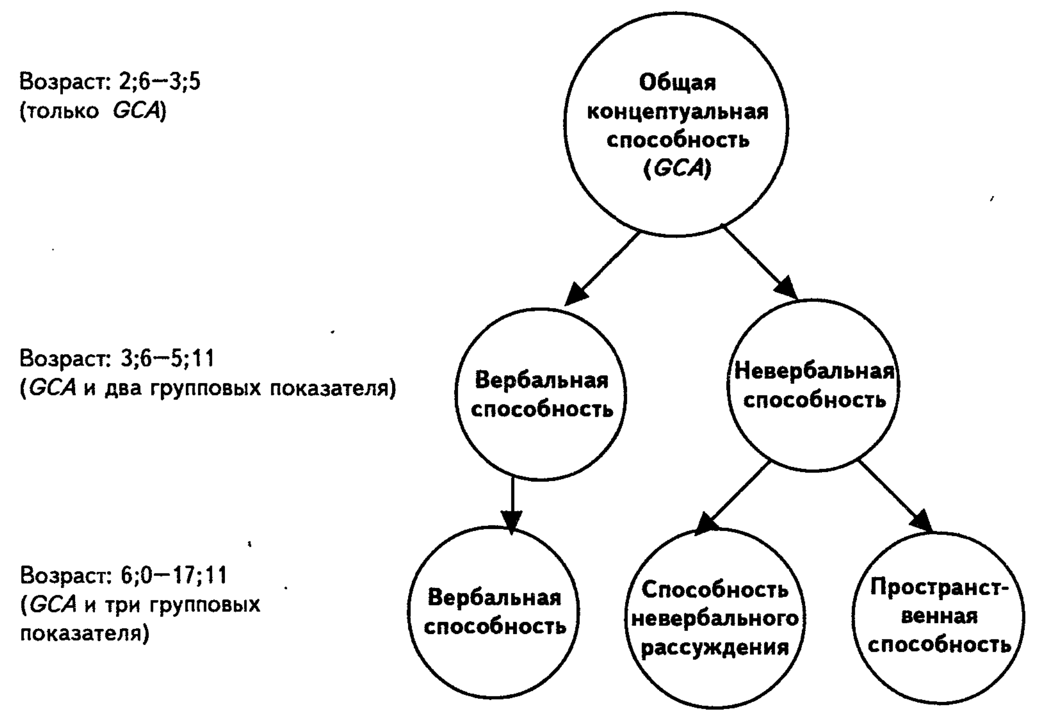


Рис. 8–6. Комбинированные показатели когнитивной батареи *DAS*
(Из Elliott, 1990b, p. 21. Copyright © 1990 by The Psychological Corporation. Воспроизведено с разрешения издателя)

к показателю *GCA*, дает еще два групповых показателя (*cluster scores*): показатели Вербальной и Невербальной способности (*Verbal and Nonverbal Ability*). Для всех испытуемых школьного возраста (от 6;0 до 17;11) можно получить три групповых показателя: Вербальной способности (*Verbal Ability*), Способности невербального рассуждения (*Nonverbal Reasoning Ability*) и Пространственной способности (*Spatial Ability*). Кроме того, посредством экстраполяций отношений между «сырыми» результатами и показателями *GCA* в разных возрастах можно расширить использование норм *GCA* до уровней выполнения существенно ниже обычных норм. Это было предусмотрено с целью облегчить оценку лиц с сильной задержкой развития, которые по всей вероятности не были представлены в нормативной выборке.

Надежность и валидность. Показатели надежности *DAS* при сравнении с таковыми у других тестов интеллекта выглядят в благоприятном свете. Коэффициенты ретестовой надежности *GCA* и групповых показателей, при интервалах между тестированиями от 2 до 7 недель, колеблются от 0,79 до 0,94. Что касается субтестов, то сопоставимые оценки ретестовой надежности варьируют здесь от 0,38 до 0,94, с подавляющим большинством, попадающим в интервал от 0,60 до 0,90. Индексы надежности как внешней согласованности оценок (*Interrater reliabilities*) для субтестов со свободными ответами, подсчет баллов в которых в значительной степени опирается на субъективные суждения,[1] группируются около 0,95.

Надежность *DAS* в плане внутренней согласованности оценивалась посредством применения теории «задание—ответ» (*IRT*). Эта процедура позволяет вычислять точные значения надежности и ошибок измерения, соответствующие каждому возможному показателю по субтесту. Полученные результаты, широко варьирующие от края к краю спектра каждой способности, подтверждают хорошо знакомую тенденцию коэффициентов надежности быть ниже для лиц на краях распределения показателя, чем для лиц, группирующихся в центре. Что касается *DAS*, оценки надежности как внутренней согласованности, вычисленные по всем предусмотренным возрастным уровням, колеблются для субтестов от 0,66 до 0,95, для групповых показателей — от 0,86 до 0,94, и для показателей *GCA* — от 0,89 до 0,96. В руководстве по *DAS*, кроме того, проводятся некоторые сравнения коэффициентов внутренней согласованности, полученных на основе *IRT* и традиционным способом. В этих случаях имеет место близкое соответствие значений коэффициентов, найденных обоими методами.

Валидность *DAS* исследовались и с внутренней, и с внешней точек зрения. Что касается внутренней валидности, для установления структуры показателей DAS (см. рис. 8–6) применялся разведочный и подтверждающий факторный анализ. Оба типа анализа дали весьма близкие результаты, которые, в целом, могут служить еще одним подтверждением ранее установленного факта прогрессирующей дифференциации способностей с возрастом (Anastasi, 1970). Однофакторная модель, включающая четыре основных субтеста, используемых для получения показателя *GCA* в возрастном диапазоне от 2;6 до 3;5, лучше всего согласуется с данными детей этой возрастной группы. Для детей в возрасте от 3;6 до 5;11 лучшим оказалось двухфакторное решение. В этом возрастном диапазоне фактор невербальной способности определялся, в основном, высокими нагрузками по таким субтестам, как «Составление фигур» и

[1] Речь идет о таких субтестах, как «Определения слов», «Аналогии», «Копирование» и «Воспроизведение образцов по памяти».

«Копирование», а фактор вербальной способности — главным образом высокими нагрузками по субтестам «Вербальное понимание» и «Называние». На уровне школьного возраста (6;0–17;11) интеркорреляции между показателями основных тестов лучше всего объясняла трехфакторная модель, охватывающая три способности: вербальную, невербального рассуждения и пространственную. Показатели пяти диагностических субтестов *DAS* не включались в комбинированный показатель *GCA* или в групповые показатели. Эти субтесты состоят в основном из задач, требующих памяти и скорости обработки информации. То, что диагностические субтесты имеют незначительные нагрузки по общему фактору (*g*) и значительную величину специфической дисперсии, делает их идеально подходящими для выявления сильных и слабых сторон индивидуума.

Обширные данные по внешней валидности *DAS* описаны в руководстве к этой батарее. К главным источникам таких данных относятся: 1) корреляции между разными показателями *DAS* (включая показатели субтестов) и показателями комплексных батарей способностей, таких как шкалы Стэнфорд—Бине и Векслера; 2) корреляции показателей субтестов *DAS* с показателями других тестов специфических когнитивных способностей и академических достижений, наподобие Словарного теста в картинках Пибоди (Пересмотренная версия — *PPVT-R*) и Тестов овладения чтением Вудкока (Пересмотренная версия — *WRMT-R*), а также со школьными оценками; 3) исследования профилей показателей *DAS* для специфических популяций: одаренных, испытывающих трудности в обучении и умственно отсталых. Все эти источники данных, в общем, подтверждают иерархическую структуру *DAS*, а также сравнимость комбинированных и частных (по субтестам и тестам достижений) показателей с аналогичными мерами. Валидность диагностических субтестов в отношении выявления подгрупп детей с трудностями в обучении хотя и выглядит достаточно убедительно, требует дополнительного исследования.

Общая оценка. Как было отмечено другими авторами (Aylward, 1992; Reinehr, 1992), сложность процедур проведения и подсчета показателей *DAS* может затормозить распространение и использование этой батареи для решения прикладных задач. Кроме того, поскольку *DAS* является относительно новым и не прошедшим клинические испытания инструментом, ему еще нужно будет на деле доказать свою ценность. Дополнительное ограничение связано с предельным возрастом обследуемых (т. е. 2;6 и 17;11), для которых минимальный и максимальный уровень трудности заданий соответственно может оказаться недостаточным. Несмотря на все это, *DAS* — измерительный инструмент, отвечающий в своей группе «современному состоянию психометрии» и до сих пор непревзойденный в отношении тех возможностей и преимуществ, которые он предоставляет своим пользователям. Иерархическая структура этой батареи, многообразие охватываемых ей способностей и та надежность, с которой она позволяет их оценивать, дают пользователю беспрецедентную гибкость в работе. В частности, проводящий тестирование может выбрать из широкого ассортимента задач те, которые лучше всего подходят для целей обследования и максимально отвечают потребностям тестируемого. Еще одной отличительной особенностью *DAS* является превосходное качество методического сопровождения этой батареи в виде двух обширных руководств: *Differential Ability Scales: Administration and scoring manual* (Elliott, 1990a) и *Differential Ability Scales: Introductory and technical handbook* (Elliott, 1990b). Последнее, в особенности, освещает практически все вопросы, какие

только могут возникнуть у пользователей, и с предельной точностью, ясностью и лаконичностью сообщает множество полезных сведений. Оно должно быть исключительно полезным для будущих пользователей батареи, особенно тех, кто хочет ближе познакомиться с теоретическими и практическими достижениями в области изучения интеллекта и его измерения, которые так хорошо представлены в итоговом продукте — Дифференциальных шкалах способностей.

Система когнитивной оценки Даса—Наглиери

Еще одним важным новым инструментом для индивидуального оценивания познавательной деятельности, опубликованным в конце 1990-х гг., стала Система когнитивной оценки Даса—Наглиери (*Das-Naglieri Cognitive Assessment System* [*CAS*]). В основу этого измерительного инструмента, разрабатывавшегося более 10 лет, положена предложенная его создателями PASS-модель[1] интеллекта (J. P. Das, Naglieri, & Kirby, 1994; Naglieri, & Das, 1990, 1997a, 1997 b). В свою очередь, их модель интеллекта основана на теории функциональной организации мозга и познания, которой придерживался российский нейропсихолог А. Р. Лурия.

Входящие в *CAS* задачи предназначены для измерения базовых когнитивных функций, участвующих в научении, но, предположительно, не зависящих от школьного обучения. К ним относятся планирование, внимание, симультанная и сукцессивная обработка информации. Система использует вербальные и невербальные тесты, предъявляемые через зрительный и слуховой сенсорные каналы. Тесты на планирование предполагают оценку стратегий, применяемых обследуемым при выполнении заданий. *CAS* рассчитана на обследование лиц в возрасте от 5;0 до 17;11 и специально проектировалась для увязывания оценки с последующим вмешательством. Благодаря ее прочной теоретической и эмпирической основе, а также тщательной, крупномасштабной стандартизации, завершение работ по *CAS* с нетерпением ожидали многие пользователи. Фактически, ее пробная версия уже получила широкое освещение в печати (Lambert, 1990; Telzrow, 1990). Судя по предварительным данным о валидности CAS, можно надеяться, что этот тест станет столь же важным, сколь и новаторским инструментом для оценки когнитивного статуса.

---

[1] Аббревиатура *PASS* образована из начальных букв названий основных переменных, входящих в данную модель: *Planning* (планирование), *Attention* (Внимание), *Simultaneous and Successive processing* (симультанная и сукцессивная обработка информации). — *Примеч. науч. ред.*

9 ТЕСТЫ ДЛЯ СПЕЦИФИЧЕСКИХ ПОПУЛЯЦИЙ

Представленные в этой главе тесты включают как индивидуальные, так и групповые шкалы. Их с самого начала разрабатывали для тестирования лиц, которые не могли быть должным образом или в полной мере обследованы традиционными инструментами, такими как описанные в предыдущей главе индивидуальные шкалы или типичные групповые тесты, обсуждаемые в следующей главе. Исторически, за рассматриваемыми в данной главе видами тестов закрепились три названия: тесты действия, неязыковые или невербальные тесты.

Тесты действия (performance tests), в целом, заключаются в манипулировании предметами, причем с минимальным использованием карандаша и бумаги. *Неязыковые тесты (nonlanguage tests)* предполагают, что ни проводящему обследование, ни обследуемому не нужно пользоваться каким-либо языком. Инструкции к этим тестам могут даваться непосредственным показом или жестами, без использования устной или письменной речи. Прототипом неязыковых групповых тестов был армейский тест бета, разработанный для тестирования во время Первой мировой войны не владеющих английским или неграмотных новобранцев (Yerkes, 1921). Впоследствии, для гражданских целей были подготовлены переработанные версии этого теста. Для большинства целей тестирования нет необходимости совершенно исключать использование языка при проведении теста, так как тестируемые обычно обладают некоторым знанием общего языка (*common language*). Кроме того, короткие, простые инструкции обычно легко переводятся или даются (последовательно) на двух языках без ощутимого влияния на существо или степень трудности теста. Впрочем, ни один из этих тестов не требует от тестируемого пользоваться при выполнении заданий письменной или устной речью.

Еще одна родственная категория тестов — *невербальные тесты (nonverbal tests)*, более правильно называемые тестами, не требующими умения читать (*nonreading tests*). К этой категории относятся большинство тестов для начальной школы и дошкольников, как и тесты для неграмотных и не умеющих читать людей любого возраста. Такие тесты, выполнение которых хотя и не требует навыков чтения и письма, предполагают широкое использование устных инструкций и речевого общения со стороны тестиру-

ющего. Более того, они часто измеряют вербальное понимание, — например, знание слов и понимание предложений или коротких абзацев, — посредством использования рисуночных заданий, дополненных и сопровождаемых устными инструкциями. Поэтому, в отличие от неязыковых тестов, они не пригодны для лиц с нарушениями слуха или не говорящих на языке тестирующего.

Хотя традиционное разграничение тестов действия, неязыковых и невербальных тестов способствует уяснению целей, которым могут служить разные тесты, различия между ними утрачивали четкость по мере того, как создавалось все больше батарей, организация которых противилась разделению входящих в них тестов на эти три категории. Классическим примером является объединение в шкалах Векслера вербальных тестов и тестов действия.

В настоящей главе тесты классифицированы не по содержанию заданий или способам предъявления, а в зависимости от основных областей их применения. С этой точки зрения можно различать четыре основные категории: тесты для младенцев и дошкольников, тесты для комплексной оценки лиц с задержкой психического развития, тесты для лиц с разными нарушениями сенсорной и моторной сферы и тесты, предназначенные для использования в различных культурах или субкультурах. Однако такая классификация должна оставаться гибкой, поскольку некоторые из тестов оказались полезными более чем в одной области применения. Это особенно справедливо по отношению к некоторым инструментам, разработанным первоначально для кросс-культурного тестирования, а в настоящее время чаще применяемым при клиническом обследовании.[1]

И последнее, хотя некоторые из тестов, рассматриваемых в данной главе, разрабатывали как групповые, их часто проводят индивидуально. Небольшая их часть широко используется при клиническом тестировании как дополняющие тесты интеллекта общего типа и тем самым обеспечивающие более полную картину интеллектуальной деятельности индивидуума. Ряд таких тестов, позволяя вести при индивидуальном тестировании определенного типа качественные наблюдения, требует значительного опыта клинических исследований для детальной интерпретации выполнения теста. В целом, все они ближе к индивидуальным тестам, рассмотренным в главе 8, чем к групповым тестам, обзору которых посвящена глава 10.

Тестирование младенцев и дошкольников

Все тесты, предназначенные для младенцев и дошкольников, требуют индивидуального предъявления. Некоторых детей, посещающих детский сад, можно объединять в небольшие группы и исследовать с помощью тестов, разработанных для учащихся начальных классов. Однако, в общем, групповые тесты непригодны для детей, не достигших школьного возраста. Большинство тестов, созданных для детей младше 6 лет, это либо тесты действия, либо устные тесты. Лишь немногие из них предполагают элементарные действия с карандашом и бумагой.

[1] Что касается дополнительной информации, оценок и ссылок на литературу, относящихся ко многим типам тестов, примеры которых приводятся в этой главе, см. Sattler (1988, chaps. 12, 14, and 15).

Принято подразделять первые 5 лет жизни на период младенчества и дошкольный период. Первый продолжается от рождения до, приблизительно, 18 мес, второй — от 18 до 60 мес. Необходимо отметить, что при проведении тестирования младенец должен либо лежать, либо находиться на коленях у взрослого, либо удерживаться взрослым в каком-то ином положении, что можно увидеть чуть позже на иллюстрациях к этой главе. Речь мало используется как средство инструктажа, хотя уровень овладения языком самого ребенка служит источником релевантных данных. Многие тесты имеют дело с сенсомоторным развитием: исследуются способности младенца поднимать голову, переворачиваться, дотягиваться до предметов и схватывать их, следить глазами за движущимся объектом. С другой стороны, дети дошкольного возраста уже могут ходить, сидеть за столом, использовать руки для манипулирования тестовыми материалами и общаться с помощью языка. В этом возрасте дети в большей степени реагируют на проводящего тестирование как на личность, тогда как для младенца он служит, главным образом, средством обеспечения стимульными объектами. Тестирование дошкольников — это в значительной степени межличностный процесс — особенность, расширяющая как возможности, так и трудности тестовой ситуации.

Корректное психологическое обследование маленьких детей требует охвата широкого спектра поведения, включая социальные и эмоциональные черты наряду с моторными, речевыми и другими когнитивными способностями. Кроме этого, наблюдается растущее признание необходимости учитывать при оценке детей характер окружения конкретного ребенка (Vazquez Nutall, Romero, & Kalesnik, 1992). Эта экологическая ориентация нашла отражение в некоторых инструментах, обсуждаемых в этой главе. В данном разделе рассматриваются типичные шкалы, предназначенные для использования в младенчестве и раннем детстве и представляющие многообразие подходов. Пересмотренная шкала интеллекта Векслера для дошкольников и младших школьников также принадлежит к этой категории, хотя и освещалась в главе 8, чтобы не разрывать обсуждение шкал Векслера. Шкала Стэнфорд—Бине, Оценочная батарея Кауфмана для детей и Дифференциальные шкалы способностей, которые также рассмотрены в главе 8, используют и для оценки детей дошкольного возраста, поскольку все они охватывают период от 2 до 6 лет в добавление к более старшим возрастам.

Исторические корни тестирования младенцев и дошкольников. Одна из самых ранних систематических попыток понять развитие нормальных младенцев и дошкольников была предпринята в серии лонгитюдных исследований Арнольдом Гезеллом и его коллегами по Йельскому университету (Ames, 1989). Эти исследования, охватившие в совокупности четыре десятилетия, привели к подготовке Таблиц развития Гезелла (*Gesell Developmental Schedules*), первая публикация которых (Gesell et al., 1940) представляла пионерскую попытку снабдить всех заинтересованных лиц систематическим, эмпирически обоснованным методом оценивания развития поведения маленьких детей. По большей части, данные для этих таблиц были получены посредством прямого наблюдения за реагированием детей на обычные игрушки и другие стимульные объекты и дополнены информацией, предоставленной родителями или воспитателями. На протяжении многих лет Таблицы Гезелла широко использовались психологами и педиатрами как в исследованиях, так и в практической работе, а после их пересмотра и обновления другими исследователями и сейчас применяются некоторыми в качестве дополнения в медицинских обследованиях, особенно для выявления неврологических дефектов и органически обусловленных отклонений в поведе-

нии на начальных этапах жизни.[1] Несмотря на то что почти во всех клинических областях применения Таблицы Гезелла были вытеснены более новыми и более тонкими в психометрическом отношении инструментами, задания и процедуры, впервые испробованные Гезеллом и его коллегами, были включены в большинство других шкал возрастного развития, предназначенных для младенческого уровня.

В период с 1960-х по 1990-е гг. наблюдался рост интереса к тестам для младенцев и дошкольников. Одним из ранних факторов, способствующих этому повышению интереса, было быстрое распространение образовательных программ для детей с задержками психического развития; другим фактором стало широкое развитие дошкольных программ компенсаторного обучения для детей, поставленных в невыгодное положение культурными барьерами. Позднее был принят целый ряд законодательных актов, способствовавших раннему выявлению и коррекции всех видов физических и психических дефектов у дошкольников и младенцев. Некоторые из этих законов (например, P. L. 99–457) являются поправками или дополнениями к Закону об образовании для всех отсталых детей (P. L. 94–142), который более подробно обсуждается чуть позже в этой же главе. Во всяком случае, под давлением этих практических нужд стали быстро появляться новые тесты и публикации и было проведено значительное число исследований, связанных с новаторскими подходами к оценке уровня развития детей.[2]

Стандартизованные тесты развития в раннем детстве

Шкалы развития младенцев Бейли. Наиболее разработанным тестом для самых ранних возрастных уровней являются Шкалы развития младенцев Бейли (*Bayley Scales of Infant Development*), иллюстрации из руководства к которым приведены на рис. 9–1. Эти шкалы, включающие в себя ряд заданий из Таблиц Гезелла и других тестов для младенцев и дошкольников, явились результатом многолетних научных изысканий Н. Бейли и ее коллег по университету в Беркли, включая лонгитюдные исследования в рамках проекта *Berkeley Growth Study*. В настоящее время пользователям доступна вторая редакция шкал Бейли (Bayley-II — Bayley, 1993).

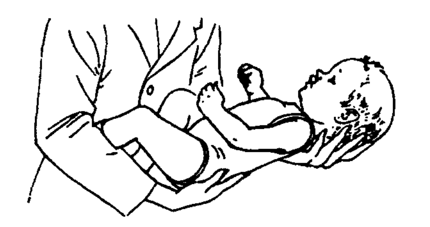
Шкалы Бейли-II предусматривают три дополняющих друг друга инструмента для оценки уровня развития ребенка в возрасте от 1 мес. до 3,5 лет: *Умственную шкалу (Mental Scale)*, *Моторную шкалу (Motor Scale)* и *Шкалу оценки поведения (Behavior Rating Scale)*. Умственная шкала позволяет проводить выборочные замеры таких функций, как острота зрения и слуха, сенсорное и перцептивное различение, память, научение, решение задач (*problem solving*), вокализация, зачатки вербального общения и элементарное абстрактное мышление. Моторная шкала служит для измерения грубых

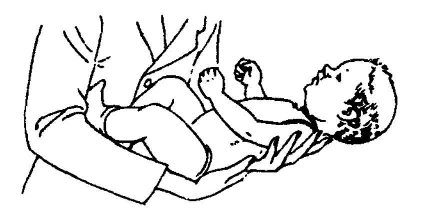
---

[1] Что касается самого свежего руководства к пересмотренной версии оригинальных Таблиц Гезелла, см. Knobloch, Stevens, & Malone (1980). Есть несколько других тестов, в названии которых используется имя Гезелла, но ни один из них не охватывает период младенчества. (См. *TIP*-IV, где помещен список всех этих тестов, имеющихся в наличии в настоящее время, и 9-й выпуск *MMY* с критическими обзорами некоторых из них).

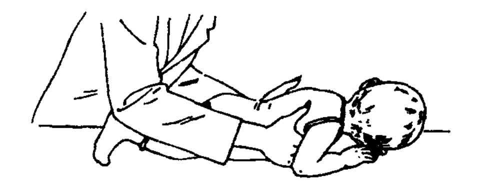
[2] Краткое, но информативное изложение истории психологического оценивания детей дошкольного возраста можно найти в работе M. F. Kelley, & Surbeck (1991). По поводу других важных сведений о тестировании младенцев и дошкольников см. Aylward (1994), Bracken (1991 b), Culbertson, & Willis (1993), Kamphaus (1993), C. R. Reynolds, & Kamphaus (1990a), Vazquez Nutall, Romero, & Kalesnik (1992).

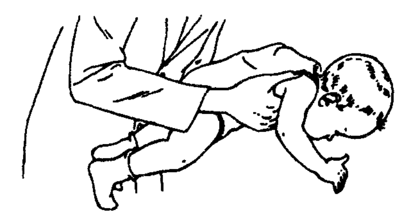
Задание 8. Поднимает голову — Поддерживание в положении на спине.





Задание 14. Корректирует положение головы при поддерживании на весу животом вниз.





Задание 33. Подтягивается в сидячее положение.



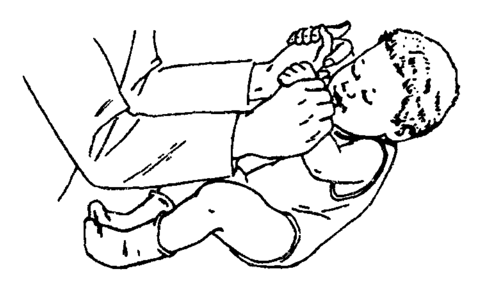


Рис. 9-1. Тестирование младенца: иллюстрации из руководства к Шкалам развития младенцев Бейли — Вторая редакция

(С упрощениями из Bayley, 1993, p. 143, 145, 150. Copyright © 1993 by The Psychological Corporation. Воспроизведено с разрешения)

моторных навыков, таких как умение сидеть, стоять, ходить и подниматься по ступенькам, а также навыков манипуляторной деятельности рук и пальцев; помимо этого, шкала включает задания для оценки сенсорной и перцептивно-моторной интеграции. В младенческом возрасте локомоторное и манипуляторное развитие играет важную роль во взаимодействии ребенка с окружающей средой и, следовательно, в развитии его умственных процессов. Оценочная шкала поведения предназначена для оценки различных аспектов развития личности ребенка, таких как эмоциональное и социальное поведение, объем внимания, уровень возбуждения (*arousal*), настойчивость и целеустремленность. Она содержит 5-балльную систему оценок для каждого задания и набор дескрипторов (или отличительных признаков) для каждого оцениваемого вида поведения. Оценочную шкалу поведения тестирующий заполняет после проведения двух других частей теста на основе сведений, полученных от ухаживающего за ребенком взрослого, и с учетом собственных впечатлений.

Шкалы Бейли выделяются среди других тестов для младенцев техническим качеством процедур конструирования заданий. Нормы для этих шкал были установлены на выборке объемом 1700 детей, по 50 девочек и 50 мальчиков в каждой из 17 возрастных групп от 1 до 42 мес. Выборка стандартизации комплектовалась таким образом, чтобы дать репрезентативный срез населения США с учетом таких характеристик, как раса / этническая группа, основные географические районы проживания и образование родителей. В нее включались только нормальные, родившиеся в срок (не раньше 36 и не позже 42 нед. беременности) дети, не имевшие сколько-нибудь серьезных медицинских осложнений и не подвергавшиеся специальному лечению по поводу психических, соматических или поведенческих проблем. Умственная и Моторная шкалы дают отдельные индексы возрастного развития, выраженные в виде нормализованных стандартных показателей со средним, равным 100, и *SD* = 15. Эти индексы вычисляются в рамках возрастной группы, в которую попадает ребенок. Возрастные группы образуются с месячным интервалом до возраста 36 мес. и с 3-месячным интервалом для более старших возрастов. Шкала оценки поведения дает процентильные показатели, которые, в свою очередь, распределяются по трем категориям: «Неоптимальный» (*Non-Optimal*), «Сомнительный» (*Questionable*) и «В границах нормы» (*Within Normal Limits*). По результатам недавно проведенного сравнительного анализа нескольких шкал для оценки детей дошкольного возраста шкала Бейли-II была признана одним из двух тестов, отвечающих стандартам технической пригодности по большинству критериев[1] (D. P. Flanagan, & Alfonso, 1995).

Бейли отмечала, что ее шкалы, как и все тесты для младенцев, следует использовать преимущественно для оценки текущего уровня развития, а не для предсказания последующих уровней способностей. На развитие способностей в столь раннем возрасте оказывает влияние такое множество промежуточных факторов, что предсказание на длительный период времени оказывается малоценным, в чем можно убедиться на основании данных, приведенных в главе 12.[2]

Со времени издания оригинальных шкал Бейли центр интересов в области тестирования развития младенцев переместился от оценивания нормальных детей раннего возраста к оценке детей с потенциальными или актуальными задержками развития. Хотя полезность шкал Бейли-II для клинических исследований далеко не исчерпана, в своем нынешнем виде эти шкалы уже включают задания, отобранные по критерию максимальной различительной способности в отношении нормальных и клинических выборок, а в руководствах к ним приводятся данные о специфических клинических популяциях. Таким образом, они должны быть полезными в обнаружении сенсорных и неврологических дефектов, эмоциональных нарушений и обусловленного средой дефицитарного развития. Кроме того, Айлвордом (Aylward, 1995) был подготовлен Скрининг-тест Бейли психоневрологического развития младенцев (*Bayley Infant Neurodevelopmental Screener* [*BINS*]), предназначенный для быстрой оценки психоневрологического статуса младенцев в возрасте от 3 до 24 мес. на основе использования комбинаций из 11–13 заданий шкалы Бейли-II и других неврологических тестов.

---

[1] Другим оказалась Пересмотренная психопедагогическая батарея Вудкока—Джонсона: Тесты познавательной способности (Woodcock, & Johnson, 1989, 1990).

[2] Обсуждение областей использования и ограничений тестов интеллекта младенцев см. в Goodman (1990). Серия статей о психометрических свойствах оригинальных шкал Бейли публикуется в Rovee-Collier, & Lipsitt (1992).

Шкалы способностей детей Маккарти. Что касается дошкольного уровня, хорошо сконструированным инструментом являются Шкалы способностей детей Маккарти (*McCarthy Scales of Children's Abilities* [*MSCA*] — McCarthy, 1972), рассчитанные на возраст от 2,5 до 8,5 лет. Они состоят из 18 тестов, предоставляющих тестирующему богатые возможности для наблюдения подхода ребенка к разнообразным задачам и стимулам. Эти тесты сгруппированы в шесть частично перекрывающихся шкал: Вербальную, Наглядно-действенную (*Perceptual-Performance*), Количественную, Общую когнитивную (*General Cognitive*), Памяти и Моторную. Показатель Общей когнитивной шкалы, основанный на результатах 15 из 18 тестов данной батареи, наиболее близок традиционной глобальной мере интеллектуального развития. Этот Общий Когнитивный Индекс (*General Cognitive Index*, или, сокращенно, *GCI*) представляет собой нормализованный стандартный показатель, выражаемый в тех же единицах, что и традиционный *IQ* (со средним, равным 100, и *SD* = 16), и вычисляется в каждой возрастной группе (с интервалом группировки 3 мес.). При разработке *MSCA* сознательно отказались от использования термина *IQ* из-за его многочисленных дезориентирующих коннотаций. *GCI* характеризуется как показатель деятельности ребенка во время тестирования и не подразумевает ничего такого, что связано с неизменяемостью или этиологией. Показатели по пяти дополнительным шкалам основаны на тех же возрастных группах и имеют среднее, равное 50, и *SD* = 10.

За два с лишним десятилетия, прошедших со времени издания шкал Маккарти, которые оказались наиболее подходящими для когнитивной оценки маленьких детей, был накоплен обширный массив данных исследований с применением этого инструмента. Особую ценность представляют многочисленные исследования, проведенные с детьми этнических меньшинств и подробно изложенные Валенсия (Valencia, 1990), а также богатейшие данные о валидности, собранные воедино им и Брэкеном (Bracken, 1991a). Что касается клинического использования шкал Маккарти, руководство к ним, подготовленное Кауфманами (Kaufman, & Kaufman, 1977), и по сей день остается обязательным пособием. По мнению многих критиков эти шкалы, несмотря на ряд слабых сторон, являются эффективным и полезным инструментом. Психометрические характеристики шкал Маккарти отвечают, по большей части, предъявляемым к ним требованиям, особенно в середине возрастного диапазона измеряемой совокупности.

Шкалы Пиаже

Будучи пригодными для изучения детей гораздо старше дошкольного возраста, эти шкалы, сконструированные на основе теорий развития Жана Пиаже, до сих пор в основном применяли при изучении раннего детства. Все эти шкалы находятся еще в стадии экспериментирования, и лишь небольшое их число издается и доступно для приобретения. По большей части их разрабатывал Ж. Пиаже для собственных программ исследования, хотя некоторые из этих шкал пригодны и для других исследовательских целей. Главный вклад шкал Пиаже в психологическое тестирование детей состоит в обеспечении теоретической системы, обосновывающей последовательность стадий развития процессов мышления, и создании процедуры оценивания, характеризующейся гибкостью и качественной интерпретацией.

Некоторые особенности шкал Пиаже в связи с нормативной интерпретацией выполнения теста обсуждались в главе 3. По существу, шкалы Пиаже являются порядковыми в том смысле, что они предполагают единую последовательность развития через

следующие друг за другом стадии. Эти стадии, охватывающие период от младенчества до юности, получили следующие названия: сенсомоторная, дооперациональная, конкретных операций и формальных операций. К тому же шкалы Пиаже соответствуют «критериально-ориентированному» подходу, поскольку дают качественное описание того, что в действительности может делать ребенок. Задачи Пиаже нацелены на изучение длительного развития у ребенка специфических понятий или когнитивных схем,[1] а не широких черт. Что же касается применения, то основная цель шкал Пиаже — «выпытать» у ребенка объяснение наблюдаемого события и выявить причины, лежащие в основе его объяснения. Подсчет баллов обычно производится исходя из качества реакций ребенка на относительно небольшое число предъявляемых ему проблемных ситуаций, а не из количества или трудности успешно выполненных заданий. По этой причине наибольший интерес представляют как раз ошибочные представления ребенка, обнаруживающие себя в его неправильных ответах. Проводящий обследование сосредоточивает основное внимание на процессе решения задачи, а не на его результате.

Из-за крайне индивидуализированных процедур проведения тесты Пиаже особенно подходят для клинической работы. Наряду с этим они привлекают внимание педагогов, поскольку позволяют объединять тестирование и обучение. И все же наиболее часто их используют в исследованиях по психологии развития. Сами тесты можно разбить на две категории: 1) порядковые шкалы для младенческого периода и 2) задачи для оценки достижения дооперационального, конкретно-операционального и формально-операционального уровней. Существует несколько образцов каждого из этих тестов, а не так давно был опубликован обзор их использования в разнообразных исследовательских контекстах (D. Sexton, Kelley, & Surbeck, 1990). Ниже мы описываем по одному тесту каждого типа, выбранных отчасти по причине их доступности.

Порядковые шкалы психологического развития (*Ordinal Scales of Psychological Development*) были подготовлены Узгирисом и Хантом (Užgiris, & Hunt, 1975). Другое название этих шкал, предназначенных для оценки приобретения когнитивных компетенций (*cognitive competencies*) в период от 2 нед. до 2 лет, — Шкалы психологического развития младенцев. Этот возраст приблизительно соответствует периоду, который Пиаже характеризовал как сенсомоторный и внутри которого он различал шесть стадий, или уровней. Чтобы повысить чувствительность своих методик, Узгирис и Хант распределили все ответы по более чем шести уровням, число которых варьирует в разных шкалах от 7 до 14. Комплект их тестов включает шесть шкал, получивших следующие названия:

1. *Постоянство объекта (Object Permanence)*: о возникающем у ребенка представлении о независимо существующих объектах судят по зрительному слежению за объектом и стремлению отыскать объект после того, как его все более тщательно прячут.
2. *Развитие средств (Development of Means)* для достижения желанных целей во внешней среде: ребенок использует свои руки и такие средства, как бечевки, палки, подставки и т. д., чтобы достать заинтересовавшие его предметы.
3. *Подражание (Imitation)*, в том числе имитация жестов и голоса.

[1] «Схемы» — термин, обычно встречающийся в работах Ж. Пиаже и обозначающий, в сущности, структуру, в которой индивид упорядочивает поступающую сенсорную информацию.

4. *Операциональная причинность (Operational Causality):* ребенок осознает объективные причинные связи и сообразовывает с ними свои действия, как показывают его реакции — от зрительного наблюдения за собственными руками до вызывания желаемого действия со стороны человека или приведения в движение механической игрушки.
5. *Отношения объектов в пространстве (Object Relations in Space):* ребенок координирует схемы смотрения и слушания, чтобы определять местоположение объектов в пространстве, и понимает такие отношения, как емкость, равновесие, тяжесть.
6. *Развитие схем (Development of Schemata)* реагирования на объекты: ребенок реагирует на объекты рассматриванием, ощупыванием, манипулированием, выпусканием из рук, бросанием и т. д., а также используя социально поощряемые схемы обращения с конкретными предметами (например, «вождение» игрушечного автомобиля, строительство из кубиков, нанизывание бусинок, называние объектов).

Норм для этих шкал нет, но авторами собраны данные об их психометрических характеристиках, полученные в результате применения шкал к 84 младенцам, которые были детьми студентов-выпускников или сотрудников университета штата Иллинойс. Приведенные сведения о согласованности результатов тестов с данными наблюдения и данными повторного тестирования, проведенного через 48 ч, говорят, в целом, об удовлетворительности обеих этих характеристик. Также сообщается, что индексы ординальности (*indices of ordinality*), подсчитанные для каждой шкалы на основе показателей той же группы из 84 детей, являются вполне удовлетворительными.[1]

Хотя и подразумевалось, что Порядковые шкалы Узгириса и Ханта носят только предварительный характер, их широко использовали с исследовательскими целями.[2] Первоначально эти шкалы предназначались для измерения влияния специфических окружающих условий на степень и ход развития младенцев. Исследования младенцев, воспитывавшихся в разных условиях, и младенцев, участвовавших в программах вмешательства, показали, что от этих средовых условий в значительной степени зависит тот средний возраст, в котором ребенок достигает разных ступеней, определяемых по шкалам развития. Эти и другие исследования, в которых Порядковые шкалы применяют для картирования когнитивного развития нормальных и отклоняющихся от «нормы» в ту или другую сторону младенцев, разбираются в книге под редакцией авторов этих шкал (Užgiris, & Hunt, 1987). Последовательность приобретений, прослеживаемых с помощью этих шкал, касается главным образом интеракций младенца с неодушевленными предметами, рассматриваемых, в свою очередь, в качестве предшественников развития коммуникативного поведения и других адаптивных навыков (Dunst, & Gallagher, 1983; Kahn, 1987).

[1] Процедуры измерения ординальности и применение шкалограммного анализа к шкалам Пиаже достаточно спорны, и это необходимо иметь в виду при интерпретации любых сообщаемых индексах порядка, относящихся к таким шкалам (F. H. Hooper, 1973; A. C. Rosenthal, 1985).

[2] Потенциальная ценность этих шкал при проведении клинической оценки также широко признается; важным шагом в направлении признания этого потенциала стала публикация руководства и форм подсчета баллов, специально предназначенных для применения шкал Узгириса и Ханта в клиническом и педагогическом контекстах (Dunst, 1980).

Другой рассматриваемый нами образец инструментария Пиаже — «Комплект для оценки понятий: Сохранение» (*Concept Assessment Kit — Conservation* [*CAK*]) — тест, официально распространяемый издателями на тех же условиях, что и другие психологические тесты. Рассчитанный на детей от 4 до 7 лет, этот тест измеряет овладение одним из наиболее известных понятий, используемых в системе Пиаже, — понятием «сохранение». Сохранение относится к пониманию ребенком, что такие свойства объектов, как вес, объем и количество, остаются неизменными, даже если объекты меняют форму, расположение, внешний вид или другие отличительные признаки. Авторы этого теста (Goldschmid, & Bentler, 1968b) выбрали понятие «сохранение» как показатель перехода ребенка от стадии дооперационального мышления к стадии конкретных операций, происходящего, по мнению Пиаже, в возрасте 7–8 лет.

Процедура проведения всего теста одинакова. Ребенку показывают два идентичных объекта, затем тестирующий производит в одном из них определенные преобразования и спрашивает ребенка, одинаковы объекты или различны. Ребенка просят пояснить свой ответ. В каждом задании 1 балл дается за правильное суждение об эквивалентности объектов и 1 балл — за приемлемое объяснение. Например, тестирующий берет два обычных стакана с равным количеством воды (континуальное количество) или с зернами кукурузы (дискретное количество) и выливает (или высыпает) содержимое либо в плоскую тарелку, либо в несколько других стаканов, меньших по величине. В другой задаче ребенку показывают два одинаковых пластилиновых шарика и затем расплющивают один, придав ему форму блина. Ребенка спрашивают, равны ли по тяжести «шар» и «блин».

Имеются три формы теста. Формы *А* и *В* параллельны и содержат по шесть задач на сохранение: Двумерное пространство, Число, Вещество, Континуальное количество, Дискретное количество и Вес. Корреляция между показателями по этим двум формам равна 0,95. Форма *С* включает две другие задачи: Площадь и Длина, — и дает корреляции с формами *А* и *В* 0,76 и 0,74 соответственно.

Нормы были установлены на выборке стандартизации, включавшей 560 мальчиков и девочек в возрасте от 4 до 8 лет из школ, центров ухода за детьми в дневное время и центров Head Start в Лос-Анджелесе (Калифорния). Эти нормы следует рассматривать лишь как предварительные ввиду малого числа испытуемых в каждой возрастной группе и недостаточной репрезентативности выборки. Средние показатели для каждой возрастной группы обнаруживают систематическое повышение с возрастом, причем особенно резкий подъем отмечается между 6 и 8 годами, что и предсказывает теория Пиаже.

Авторами *САК* проведен многоцелевой статистический анализ, в результате которого были определены различные типы надежности (ретестовая, параллельных форм, Кьюдера—Ричардсона, а также надежность оценщика); получены оценки шкалируемости (*scalability*), или ординальности, а также факторная структура (см. также Goldschmid, & Bentler, 1968a). Результаты, хотя они и получены на относительно малых выборках, в общем, свидетельствуют об удовлетворительной надежности, подтверждают ординальность шкалы и указывают на присутствие значительного общего фактора (*common factor*) сохранения во всех задачах.

Сравнительные исследования, проведенные в семи странах, подтвердили, что тест пригоден для применения в разных культурах, дает высокие коэффициенты надежности и выявляет приблизительно одни и те же тенденции возрастного развития (Goldschmid et al., 1973). Но в разных культурах и субкультурах были обнаружены различия

в среднем возрасте овладения понятиями, — т. е. возрастная кривые могут смещаться по горизонтали на один или два года (см. также Figurelli, & Keller, 1972; Wasik, & Wasik, 1971). Было обнаружено, что тренировка в решении задач на сохранение значительно улучшает показатели (см. также Goldschmid, 1968; B. J. Zimmerman, & Rosenthal, 1974a, 1974b). В руководстве к САК приводятся внушительные данные о конструктной валидности этого теста, которые, в целом, подтверждают в своем недавнем исследовании Ф. Кэмпбелл и Рэйми (F. A. Campbell, & Ramey, 1990).

Оценка пиажетианского подхода. Споры по поводу теоретических основ и эмпирической обоснованности подхода Пиаже к когнитивному развитию продолжаются до сих пор (см., например, Inhelder, de Caprona, & Cornu-Wells, 1987; Liben, 1983; Sugarman, 1987). По-прежнему нет окончательных ответов на вопросы о значении эффектов обучения и о влиянии кросс-культурных различий на интерпретацию пиажетианских стадий развития. Главное препятствие, с которым приходится сталкиваться при идентификации стадий с помощью порядковых шкал, заключается в том, что пиажетианцы называют *декаляжем (dècalage)*,[1] или нарушениями ожидаемого порядка следования. Непрерывно растет корпус данных, подвергающих сомнению последовательность и регулярность хода интеллектуального развития. Слишком часто стадия, соответствующая результатам конкретного ребенка, изменяется вместе с изменением задачи, причем не только в тех случаях, когда для ее решения необходимы другие способы, но и тогда, когда те же способы применяются к другому содержанию (Dasen, 1977; Goodnow, 1976; Horn, 1976; McV. Hunt, 1976).

Следует также отметить, что шкалы Пиаже коррелируют в значительной степени со стандартизованными тестами интеллекта (Gottfried, & Brody, 1975; Kaufman, 1971; M. E. Sexton, 1987), и в той же мере коррелируют с учебными достижениями первоклассников, как и групповой тест интеллекта (Kaufman, & Kaufman, 1972). Такое перекрытие получило прочное подтверждение со стороны независимых исследователей, работавших с разными инструментами (Humphreys, Rich, & Davey, 1985). Эти результаты говорят о том, что несмотря на явные различия в методологии шкалы Пиаже, стандартизованные тесты интеллекта и меры учебных достижений имеют много общего. К тому же каждый из подходов вносит неповторимые и ценные элементы в общую оценку детей. Шкалами Пиаже труднее пользоваться, и они требуют существенно больше времени для обследования детей, но они дают гораздо более богатую картину того, что может делать ребенок и как он это делает, особенно когда эти шкалы используются в сочетании с критериально-ориентированными и нормативно-ориентированными мерами (D. Sexton et al., 1990).

Современные исследования умственной деятельности маленьких детей представляют собой быстро развивающуюся область. Получаемые в них эмпирические результаты способствуют пересмотру и расширению ранних концепций Пиаже (см., например, Butterworth, Harris, Leslie, & Wellman, 1991; Whiten, 1991). Фактически, в наше время существует ряд новых подходов, объединенных под названием «неопиажетианского», которые занимаются изучением проблем когнитивного развития в перспективе, определяемой различными комбинациями положений теории Пиаже и теории обработки информации (Beilin, & Pufall, 1992; Demetriou, 1988). В области

---

[1] Буквально: «расклинивание» (*unwedging*), или расхождение теоретически ожидаемого паттерна реакций.

психологической оценки некоторые исследователи-«неопиажетианцы» объединяют разнообразные динамические подходы и, используя промежуточное обучение в формализованной манере, пытаются оценить чистую умственную способность (*mental capacity*) с минимальной опорой на предыдущие знания индивидуума (Pascual-Leone, & Ijaz, 1991). Эти методики, которые пока еще носят экспериментальный характер, по расчетам их создателей должны быть применимы как к маленьким детям (в возрасте 2–3 лет), так и к представителям разных культур, социальных слоев и языковых групп.

Современные тенденции в оценивании младенцев и детей раннего возраста

В историческом плане валидность тестов интеллекта была связана главным образом с критериями возрастной дифференциации и корреляциями их показателей с результатами учебной деятельности. Что касается младенцев, адекватное продвижение вперед измерялось почти исключительно с помощью сравнения их результатов с нормами для того же возраста по широкому кругу задач, включаемых в шкалы возрастного развития, наподобие шкал Бейли. Однако усилия современного общества, направленные на раннее выявление и коррекцию дефицитарного развития детей, требуют, чтобы инструменты, предлагаемые для оценки познавательной деятельности младенцев, обладали прогнозирующей силой. Поэтому несмотря на трудности, порождаемые намерением проследить связанные с развитием изменения в интеллектуальной компетенции на разных возрастных уровнях, были возобновлены попытки создать инструменты и методики, которые бы обладали достаточной для практических целей прогностической ценностью.

Один из наиболее интересных результатов этих новых подходов заключается в создании средств измерения навыков обработки информации, таких как Тест интеллекта младенцев Фэгана (*Fagan Test of Infant Intelligence* — Fagan, 1992; Fagan & Detterman, 1992). Этот подход основан на твердо установленных данных о предпочтении младенцами новых раздражителей, которое, в свою очередь, делает возможным изучение их способности абстрагировать и сохранять информацию в памяти. Тест Фэгана, предназначенный для дифференциации нормальных детей и детей с когнитивным недоразвитием, оценивает избирательное зрительное внимание к новым раздражителям у младенцев в возрасте от 3 до 12 мес. В качестве раздражителей используют изображения человеческих лиц, а «показатель» основан на суммарном времени, уделенном новым (противопоставляемым знакомым) изображениям. На рис. 9–2 показана переносная настольная версия этого инструмента, который, как оказалось, предсказывает более позднее выполнение интеллектуальных тестов не хуже или даже лучше, чем традиционные меры интеллекта младенцев. Корреляции между показателями теста Фэгана и *IQ* в трехлетнем возрасте колеблются где-то от 0,45 до почти 0,60. Тест Фэгана находится еще в стадии разработки, но уже критиковался по ряду пунктов (см., например, Benasich, & Bejar, 1992; Goodman, 1990). Бесспорно, необходимо накопить больше данных о его клинической полезности в предсказании недоразвития отдельных когнитивных функций, включая и умственную отсталость. Тем не менее сам подход к созданию этого теста имеет твердую эмпирическую основу и полностью согласуется с данными о природе младенческого интеллекта, обсуждаемыми в главе 12.

Отмечается также растущее понимание того, что, если мы хотим повысит эффективность программ вмешательства, оценка маленьких детей должна быть настолько

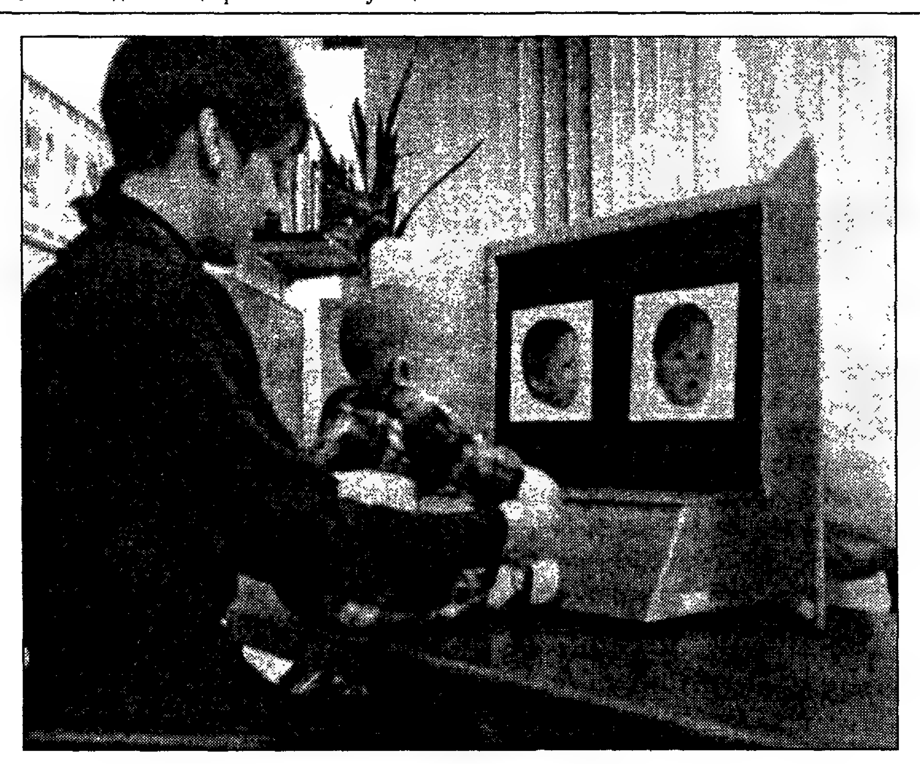


Рис. 9–2. Портативная настольная версия Теста интеллекта младенцев Фэгана *(Из Fagan & Detterman, 1992, p. 189. Copyright © 1992 by Ablex Publishing. Воспроизведено с разрешения)*

же всесторонней, насколько точной и валидной. Опора на единственный глобальный показатель, называется ли он *IQ* или индексом возрастного развития (*developmental index*), явно не отвечает большинству практических целей. Такие показатели могут служить для распределения детей по разным категориям, но они не информируют пользователей о сильных и слабых сторонах ребенка. Индивидуализированная оценка младенцев и детей, так же как и лиц более старшего возраста, требует использования комплексных методов и разнообразных источников информации, чтобы охватить все релевантные области, такие как язык, моторику и социальные навыки. В ответ на эти требования в данное время разрабатываются новые системные методы оценки, в создании которых принимают участие разные специалисты, вносящие в психологическое оценивание приемы из других дисциплин.

Система оценки возрастного развития младенцев и детей раннего возраста (*Infant-Toddler Developmental Assessment* [*IDA*]), основанная на результатах работы междисциплинарной группы специалистов по раннему детству (Provence, Erikson, Vater, & Palmeri, 1995a, 1995b, 1995c), служит примером данной тенденции. *IDA* — это по существу система методических материалов, направляющая групповую работу (*team process*) по выявлению детей (в возрасте от рождения до 3 лет) с риском задержек развития. Входящие в *IDA* материалы помогают собрать, запротоколировать, проинтерпретировать и синтезировать данные по всем линиям плана развития. Используемые процедуры включают привлечение родителей на каждом этапе обследования, терапевтический

осмотр (*health review*) и оценку развития, основанную на наблюдении и беседах с родителями и другими лицами, участвующими в уходе за ребенком. Составная часть *IDA*, имеющая наибольшее сходство с традиционными средствами оценки развития младенцев, называется Профилем развития от рождения до трех лет Провинс (*Provence Birth to Three Developmental Profile*). Хотя этот профиль предполагает использование типичных процедур проведения тестов и подсчета баллов при оценивании восьми областей развития и соотношений между ними, его показатели являются возрастно-ориентированными (*age-referenced*), т. е. основанными на документально подтвержденных вехах возрастного развития, а не на стандартизованных оценках, процентилях или других внутригрупповых сравнениях. В этом отношении педиатр Сэлли Провинс — создатель этого профиля — следовала традиции, установленной Гезеллом в его Таблицах возрастного развития. Эффективность *IDA*, как и любого другого инструмента, несомненно связана с профессиональной подготовкой и опытом применяющих эту систему специалистов; ее еще предстоит оценить в последующей исследовательской и клинической работе с *IDA*. Тем не менее *IDA* и другие подобные ей системы были задуманы как ответ на критику, направленную против сложившейся практики излишне доверять тестам интеллекта (см., например, Goodman, 1990), и, при должном воплощении замысла, могут стать крайне ценными, в практическом смысле, инструментами. Следует заметить, что использование *IDA* не препятствует проведению традиционных измерений познавательной деятельности или применению других традиционных методов, таких как оценка относительного положения ребенка в группе сверстников, когда в этом есть необходимость.

Помимо постепенного перехода к более комплексной и интегрированной оценке, существуют две другие тенденции, оказывающие значительное влияние на тестирование маленьких детей, как, впрочем, и на тестирование большинства других специфических популяций, обсуждаемых в этой главе. Первая — влияние экологической перспективы на процесс оценивания, которая требует принимать в расчет различные аспекты окружающей ребенка среды.[1] Вторая — придание все большего значения связи между оценкой и вмешательством, важность которой уже давно признавалась в клинической работе, где диагноз неразрывно связан с лечением. Потребность в создании предписывающих инструкций для учителей на основе индивидуального профиля способностей и «неспособностей» (*disabilities*) ребенка рассматривается в настоящее время как одна из важнейших в контексте раннего вмешательства и обучения (Bagnato & Neisworth, 1991; Witt, Elliott, Gresham, & Kramer, 1988).

Комплексная оценка лиц с задержкой психического развития

Тестирование умственно и физически неполноценных детей испытало в США заметный скачок развития вслед за принятием в 1975 г. Закона об образовании для всех отсталых детей (*Education for All Handicapped Children Act* — P. L. 94–142) — после внесения поправок называемого Законом об образовании для умственно и физически неполноценных лиц (*Individuals with Disabilities Education Act* [*IDEA*]) — и связанных

[1] Оценка среды обсуждается в последнем разделе этой главы.

с ним законодательных актов.[1] Реализация этого закона требует четырех основных процедур: 1) всех детей с различными видами «неспособностей» (*disabilities*) следует выявлять с помощью инструментов предварительного отсеивания; 2) выявленных таким способом детей должна оценивать группа специалистов с целью определения образовательных потребностей каждого ребенка; 3) школа должна разрабатывать индивидуализированные программы обучения, отвечающие этим потребностям; 4) каждого ребенка следует периодически оценивать в ходе обучения по разработанной для него программе. Тесты, пригодные для использования в образовательных программах, отвечающих требованиям данного закона, рассматриваются в нескольких местах нашего учебника, включая главы 8 и 17, а также в этом и двух следующих разделах настоящей главы (см. также Jacobson, & Mulick, 1996).

В руководстве по терминологии и классификации Американской ассоциации по изучению умственной отсталости (*American Association on Mental Retardation — AARM*, 1992) приводится определение, согласно которому термин «психическая задержка (*mental retardation*) относится к существенным ограничениям в текущей деятельности. Психическая задержка характеризуется интеллектуальной деятельностью на уровне значительно ниже среднего в сочетании со связанными с этим ограничениями в двух или более нижеупомянутых областях применения адаптивных навыков: общение, самообслуживание, домашняя жизнь, социальные навыки, участие в жизни общины, руководство собой (*self-direction*), здоровье и безопасность, учеба, досуг и работа» (р. 1). В руководстве также особо оговаривается, что состояние «психической задержки» выявляется до 18 лет. Это определение не отличается сколько-нибудь существенно от прежнего определения (Grossman, 1983, р. 11). Однако связанная с ним система классификации заметно изменилась, с тем чтобы приспособиться к господствующей точке зрения, согласно которой психическая задержка представляет собой не черту, а неспособность (*disability*), возникающую в результате взаимодействия ограничений индивидуума и требований окружающей среды. Тогда как прежняя система точно определяла уровни умственной отсталости — от легкой до глубокой — на основе результатов теста интеллекта, нынешняя система классифицирует уровни поддержки (*intensities of supports*), необходимой индивидууму по четырем направлениям: 1) интеллектуальная деятельность и адаптивные навыки, 2) соображения психологического/эмоционального порядка, 3) физическое здоровье / этиология и 4) учет требований среды. Определения уровней поддержки, вместе с соответствующими примерами, приведены в табл. 9–1[2].

Новое определение согласуется с прежним представлением о том, что интеллектуальное ограничение является необходимым, но недостаточным условием диагноза «психическая задержка». Для существования последней интеллектуальное ограничение должно сказываться на навыках адаптивного или совладающего поведения. За пороговый уровень выполнения теста интеллекта, используемый для разграничения

[1] Важнейшими из них являются P. L. 99–457 и P. L. 101–476, которые были приняты в 1986 и 1990 гг. соответственно. Обсуждение последствий принятия этого федерального закона и связанных с его применением судебных прецедентов для развития тестирования и оценки детей см. в Ayers, Day, & Rotatori (1990), DeMers, Fiorello, & Langer (1992), M. P. Kelly & Melton (1993), Sattler (1988, p. 767–784).

[2] В последнем издании Руководства по диагностике и статистической классификации психических расстройств (DSM-IV — 1994), подготовленном Американской психиатрической ассоциацией, все еще сохраняется определение четырех степеней умственной отсталости на основе уровней *IQ*, а именно: легкой, средней, тяжелой и глубокой.

Таблица 9-1

Определения и примеры уровней поддержки

Эпизодическая

Поддержка «по требованию». Носит эпизодический характер: человек не нуждается в постоянной поддержке или нуждается только в кратковременной поддержке в время переходов на жизненном пути (например, при потере работы или при обострении болезни). Периодическая поддержка, когда это предусмотрено, может высоко или, наоборот, низко интенсивной.

Ограниченная

Уровень поддержки, характеризуемый постоянством на протяжении некоторого, хотя бы и ограниченного периода времени, может требовать меньшего количества персонала и меньших финансовых расходов по сравнению с более глубокими уровнями поддержки (например, обучение при поступлении на работу или поддержка при налаживании взрослой жизни после окончания школы).

Экстенсивная

Поддержка оказывается регулярно (например, ежедневно), по крайней мере, в некоторых условиях (таких, как место работы или дом), и не ограничена во времени (например, долговременная психологическая поддержка и долговременная помощь по ведению домашнего хозяйства).

Полная

Поддержка характеризуется постоянством и высокой интенсивностью, оказывается во всех условиях жизнедеятельности и направлена, по существу, на поддержание жизни. В типичных случаях полная поддержка требует большего количества персонала и более глубокого вмешательства в жизнь человека, чем экстенсивная или ограниченная поддержка.

(Из AAMR, 1992, р. 26. Copyright © 1992 by the American Association on Mental Retardation. Воспроизводится с разрешения.)

нормы и психической задержки, принимают показатель от 70 до 75 единиц по шкале со средним 100 и SD = 15, что *примерно* соответствует двум или более стандартным отклонениям ниже среднего; тем самым учитывается ошибка измерения и подчеркивается факт отсутствия резкой разделительной линии между «умственной отсталостью» и «нормой». Отказ от использования уровней (или степеней) отсталости и возросший акцент на адаптивных навыках и требованиях среды имеют целью сосредоточить внимание на уникальном сочетании сильных и слабых сторон у конкретного человека и на возможностях роста. Ревизии *AAMR* в этой области вызвали дискуссию. Некоторые критики обвинили *AAMR* в том, что ее новые инструкции и нормативы расплывчаты, не поддаются надежной оценке и будут способствовать увеличению доли населения, имеющего право на получение специальных образовательных услуг в школах (Gresham, MacMillan, & Siperstein, 1995; MacMillan, Gresham, & Siperstein, 1993; Matson, 1995). Противная сторона заявила, что эти обвинения беспочвенны (Reiss, 1994). В любом случае, принятие иных стандартов скорее всего отразится как на манере проведения обследования, так и на его результатах; однако сейчас было бы преждевременно оценивать все последствия этих ревизий.

Помимо индивидуальных тестов интеллекта, наподобие описанных в главе 8, программы оценки лиц с задержкой психического развития обычно включают средства измерения адаптивного поведения в ситуациях повседневной жизни.[1] Прототипом

[1] Следует заметить, что не все из главных индивидуальных шкал интеллекта одинаково хорошо работают при применении к лицам с задержками психического развития (см., например, Spruill, 1991).

шкал для оценки адаптивного поведения является Вайнлендская шкала социальной зрелости (*Vineland Social Maturity Scale*), разработанная в 1930-х гг. директором Вайнлендской исправительной школы Эдгаром Доллом (Doll, 1935/1965). Опираясь на результаты собственных наблюдений за различиями в поведении умственно отсталых воспитанников школы, Долл создал стандартизованную регистрационную форму для оценки возрастного уровня развития индивидуума в том, что касается способности самообслуживания, удовлетворения практических нужд и выполнения обязанностей в повседневной жизни. Самая последняя переработка этой шкалы — Вайнлендские шкалы адаптивного поведения (*Vineland Adaptive Behavior Scales* [*VABS*] — P. L. Harrison, 1985; Sparrow, Balla, & Cicchetti, 1984a, 1984b) — доступна в трех версиях, которые можно использовать независимо или в сочетании. Две из этих шкал представляют собой схемы слабоструктурированных интервью, предназначенных для сбора информации в процессе беседы с родителем или воспитателем. Одна — Обзорная форма (*Survey Form*) из 297 вопросов, более всех других напоминающая первые варианты Вайнлендской шкалы. Другая — состоящая из 577 вопросов Расширенная форма (*Expanded Form*), которая к тому же обеспечивает систематическую основу для подготовки индивидуализированных обучающих или терапевтических программ. Обе применимы начиная с рождения и до 18 лет (и могут распространяться на взрослых с выраженным снижением интеллекта). Третья версия — Педагогическая (*Classroom Edition*) — представляет собой вопросник из 244 пунктов, заполняемый учителем, и рассчитана на возраст от 3 до 12 лет. Корреляции между показателями Педагогической и Обзорной форм колеблются от 0,31 до 0,54, что говорит о недопустимости использования этих форм как взаимозаменяемых.

Все версии Вайнлендских шкал нацелены на оценку того, что индивидуум обычно и охотно *делает*, а не того, что он *способен* сделать. Все их вопросы сгруппированы по четырем главным областям адаптивного поведения, показанным на рис. 9–2 вместе с подобластями и краткими характеристиками охватываемого поведения. Обе формы интервью — Обзорная и Расширенная — включают, кроме того, дополнительный набор вопросов (32), касающихся дезадаптивного или нежелательного поведения, которое может мешать нормальной жизнедеятельности. Все версии содержат в комплекте детально разработанные формы заключения по результатам обследования, предоставляемого родителям.

Обе версии шкал в форме интервью были стандартизованы на национальной выборке объемом 3000 человек в возрасте от рождения до 18 лет 11 мес., стратифицированной на основании данных о переписи населения США 1980 г. по полу, этнической принадлежности, величине населенного пункта, географическому району и образованию родителей. Кроме того, были установлены нормы для специальных групп, включая выборки живущих дома и в специальных учреждениях взрослых с диагнозом умственной отсталости, а также выборки детей с эмоциональными нарушениями и детей с ослабленным зрением и слухом, живущих в домашних условиях. Педагогическая версия шкалы была стандартизована на выборке объемом около 3000 детей в возрасте от 3 лет до 12 лет 11 мес., сформированной из учащихся школ в 38 штатах и стратифицированной по тем же критериям, которые использовались при стандартизации других версий.

Все три формы обеспечивают получение показателей по четырем областям и Комплексного показателя адаптивного поведения в виде стандартных показателей со средним, равным 100, и $SD = 15$. Указываются также диапазоны ошибок (основанные

Таблица 9-2

Содержание Вайнлендских шкал адаптивного поведения

Области и подобласти	Характеристика
Коммуникация	
Рецептивная	Что обследуемый понимает
Экспрессивная	Что обследуемый говорит
Письменная	Что обследуемый читает и пишет
Навыки повседневной жизни	
Личные	Как обследуемый ест, одевается и пользуется средствами личной гигиены
Связанные с работой по дому	Какие виды домашнего труда выполняет обследуемый
Общественные	Как обследуемый расходует время, деньги, пользуется телефоном и трудовыми навыками
Социализация	
Межличностные отношения	Как обследуемый взаимодействует с другими
Игра и проведение досуга	В какие игры играет обследуемый и как использует свободное время
Умение уживаться с другими	Насколько обследуемый проявляет ответственность и чувствительность к потребностям других людей
Двигательные навыки	
Грубая моторика	Как передвигается обследуемый и насколько у него развита координация рук и/или ног
Тонкая моторика	Как обследуемый использует руки и пальцы при обращении с предметами
Комплексный показатель адаптивного поведения	Вычисляется на основе суммирования показателей по четырем указанным выше областям
*Дезадаптивное поведение**	Нежелательное поведение, которое может мешать приспособительной деятельности

* Только при использовании дополнительного набора вопросов в формах Обзорного и Расширенного интервью

(С упрощениями из Sparrow, Balla, & Cicchetti, 1984a, p. 3. Copyright © 1984 by American Guidance Service, Inc. Воспроизводится с разрешения издателя. All rights reserved.)

на *SEM*) для пяти различных доверительных уровней (от 68 % до 99 %). Кроме того, предусмотрен перевод тех же суммарных показателей в процентили, станайны, возрастные эквиваленты и уровни адаптации (качественные описательные категории). Что касается показателей по подобластям адаптивного поведения, то здесь результаты могут выражаться в виде уровней адаптации или возрастных эквивалентов; дезадаптивное поведение оценивается только качественно, в виде уровней дезадаптации. Можно воспользоваться дополнительными нормами для получения процентильных и качественных (уровневых) показателей в каждой из специальных групп. Имеется программное обеспечение для преобразования первичных показателей в производные и для анализа профиля.

Для всех версий *VABS* средние коэффициенты надежности как внутренней согласованности частных (по областям) и комплексного показателей по большей части превышают 0,90. Понятно, что надежность этого типа ниже для подобластей, и ее коэффициенты широко варьируют в зависимости от возрастного уровня и содержательной области. Тем не менее, и для подобластей средние коэффициенты надежности в основном больше 0,70–0,80. Данные по ретестовой надежности и надежности оценщика говорят о хорошей устойчивости показателей в небольших временных интервалах и удовлетворительной согласованности между результатами двух интервьюеров, применявших шкалу к одним и тем же респондентам.

Сводки некоторых типов данных, приведенных в руководствах к трем формам *VABS*, вносят вклад в конструктную валидность этих шкал. В известной степени, валидность была заложена в этот инструмент с самого начала благодаря формулированию конструктов адаптивного поведения, которые направляли подготовку и отбор вопросов. Данные эмпирической валидизации получены в результате анализа выборок стандартизации, а также представлены в публикациях независимых исследователей. Они включают данные о тенденциях возрастного развития в областях и подобластях, охватываемых этими шкалами; результаты факторного анализа показателей по областям и подобластям шкал; сравнения профилей показателей, полученных на выборках умственно отсталых и имеющих эмоциональные и сенсорные дефекты, обследованных с целью установления дополнительных норм; корреляции с другими инвентарями адаптивного поведения и такими тестами способностей, как *WISC-R*, *K-ABC* и Словарный тест в картинках Пибоди.

В общем, процедуры, используемые при разработке и оценке Вайнлендских шкал, отличались высоким техническим качеством и достаточно полно и ясно описаны в руководствах к ним. Они отражают достижения в процессе конструирования тестов, которые появились со времени публикации ранних редакций этих шкал. Однако практическая эффективность данного инструмента зависит от полного знания его психометрических характеристик, описанных в руководстве и посвященных ему публикациях, накопившихся к настоящему времени.[1]

Как уже отмечалось, имела место широкая заинтересованность в использовании результатов оценки для проектирования и подбора подходящих программ обучения для лиц с психической задержкой. Этот интерес, в свою очередь, привел к разработке все большего количества шкал для измерения адаптивного поведения.[2] Один из примеров — Шкалы адаптивного поведения, разработанные *AAMR* и предназначенные для тех же целей, что и Вайнлендские шкалы. *AAMR* шкала адаптивного поведения (для живущих дома и в специальных учреждениях) — Вторая редакция (*AAMR Adaptive Behavior Scale — Residential and Community, Second Edition* [*ABS-RC:2*] — Nihira, Leland, & Lambert, 1993) — была стандартизована на более 4000 умственно неполноценных (в результате задержки развития) взрослых, проживавших в специальных учреждениях или дома. Она позволяет получать показатели по 18 областям, 10 из

[1] См., например, Middleton, Keene, & Brown (1990), Poth, & Barnett (1988), Raggio, & Massingale (1990), Schatz & Hamdan-Allen (1995), Silverstein (1986). Дополнительную характеристику и независимые оценки трех форм Вайнлендских шкал можно найти в статьях I. A. Campbell (1985) и C. R. Reynolds (1986).

[2] Обзоры многих из этих шкал см. в Fox, & Meyer (1990), Knoff (1992) и Sattler (1988, chap. 15).

которых относятся к навыкам управления своим поведением и 8 — к социальному поведению, включая различные типы дезадаптивных моделей. С другой стороны, *AAMR* шкала адаптивного поведения (для учащихся) — Вторая редакция (*AAMR Adaptive Behavior Scale — School, Second Edition* [*ABS-S:2*] — Lambert, Nihira, & Leland, 1993) — была нормирована на детях с задержками и без задержек психического развития в возрасте от 3 до 18 лет.

Еще одна область, требующая оценки у лиц с психической задержкой, — моторное развитие (обследуемое также с помощью шкал для младенцев). Прототипом инструментария, используемого для этой цели, являются Тесты двигательных умений Озерецкого, впервые опубликованные в России в 1923 г. Другое применение тесты Озерецкого находят в тестировании детей с нарушениями моторики, минимальной мозговой дисфункцией или трудностями в обучении, особенно в связи с организацией программ индивидуализированного обучения. Современная пересмотренная версия шкал Озерецкого — Тест двигательных умений Брунинкса—Озерецкого (*Bruininks—Oseretsky Test of Motor Proficiency* — Bruininks, 1978).

Полная батарея, требующая для проведения обследования от 45 до 60 мин, содержит 46 заданий, сгруппированных в 8 субтестов. Батарея дает три показателя: Комплексный показатель грубой моторики (*Gross Motor Composite*), служащий мерой деятельности крупных мышц плечевого пояса, туловища и ног; Комплексный показатель тонкой моторики (*Fine Motor Composite*), оценивающий деятельность мелких мышц пальцев, кистей и предплечий, и Комплексный показателей полной батареи (*Total Battery Composite*).

Кроме того, имеется краткая форма этого теста, состоящая из 14 заданий, требующая от 15 до 20 мин на проведение и дающая только один показатель: индекс общего моторного развития (*index of general motor proficiency*). Результаты выполнения теста могут выражаться в стандартных показателях для определенных возрастных групп, процентилях и станайнах. Выполнение каждого субтеста можно также представить в виде возрастных эквивалентов. Батарея была стандартизована на выборке объемом 765 детей в возрасте от 4,5 до 14,5 лет, репрезентативно отражающей данную часть населения США. Ретестовая надежность комплексных показателей в интервалах от 7 до 12 дней составляет величину порядка 0,80. Валидность батареи исследовалась несколькими способами, включая факторный анализ показателей, изучение возрастной дифференциации и сравнение нормальных детей с детьми, имеющими психическую задержку и трудности в обучении.

Одна из самых серьезных трудностей, связанных с оценкой психической задержки, заключается в необходимости разграничения между этим состоянием (*mental retardation*) и замедлением развития (*developmental delays*), особенно в младенчестве и раннем детстве. Дело не только в том, что оценка когнитивного развития в этот период менее надежна, чем в других возрастах, но и в том, что наблюдаемое отставание в познавательной деятельности может быть следствием разнообразных состояний (Hodapp, Burack, & Zigler, 1990).

Главными среди факторов, оказывающих негативное воздействие на уровень интеллектуальной деятельности и адаптивных навыков, являются сенсорные и моторные нарушения, а также неблагоприятная домашняя среда. Оставшаяся часть этой главы посвящена обсуждению вопросов, связанных с этими факторами, которые могут действовать и по отдельности, и в сочетании.

Тестирование лиц с физическими недостатками

Несмотря на то что проблемами, которые ставило тестирование лиц с физическими недостатками, специалисты занимались не одно десятилетие, особое внимание к ним было стимулировано законами, принятыми после 1970 г. Обеспечение подходящего образования для всех детей с физическими недостатками предусматривается уже упоминавшимся Законом об образовании для умственно и физических неполноценных лиц. На более широком уровне общие положения Закона о гражданских правах, распространяемые на другие меньшинства, были расширены с целью охватить лиц с физическими недостатками, сначала благодаря параграфу 504 Закона о реабилитации инвалидов (1973), а позднее — благодаря Закону об инвалидах-американцах (*Americans with Disabilities Act*) от 1990 г. (*ADA* — P.L. 101–336).[1] Эти законы запрещают дискриминацию в областях: 1) найма на работу; 2) доступности физических удобств и технического оборудования; 3) получения дополнительного (*postsecondary*) образования после окончания школы и 4) услуг медицинского и социально-бытового характера. Закон об инвалидах-американцах усиливает уже имевшиеся в американских законах права таких лиц и распространяет их на организации в частном секторе.

Тестирование детей с физическими недостатками в раннем возрасте представляется особенно важным в связи с необходимостью обеспечить им подходящий образовательный опыт с самого начала. Такой подход способствует предотвращению накопленных дефектов обучения, которые могли бы усилить воздействия конкретного недостатка на интеллектуальное развитие.[2] В любом возрасте тестирование лиц с физическими недостатками сопряжено с рядом специфических проблем в отношении проведения теста и правильной интерпретации его результатов. И по сей день основные пути решения этих проблем заключаются в 1) изменении способа тестирования, лимитов времени и содержания существующих тестов и 2) индивидуализированной клинической оценке, которая объединяет тестовые показатели с данными, получаемыми из других источников: биографических сведений, интервьюирования и оценок со стороны хорошо осведомленных наблюдателей, например учителей (*AERA*, *APA*, *NCME*, 1985, chap. 13; Bailey, & Wolery, 1989; Barnett, 1983; Culbertson, & Willis, 1993; Eyde, Nester, Heaton, & Nelson, 1994; Scarpati, 1991; Sherman, & Robinson, 1982).

Попытки установить отдельные нормы для людей с конкретными физическими недостатками или сконструировать тесты специально для таких групп обычно наталкиваются на препятствие в виде малого числа доступных для обследования лиц. Это ограничение связаны главным образом с малой долей таких лиц и множественными дефектами, а также с использованием тестов в определенных контекстах, — таких как прием в аспирантуру и в профессиональные школы, — предполагающих специально отобранные выборки. Однако все эти трудности не могут остановить исследования деятельности лиц с различными физическими недостатками, в которых им предъявляются стандартные или специально адаптированные версии разных тестов.

---

[1] Анализ последствий Закона об инвалидах-американцах для психологического тестирования см. в Nester (1994). Поднимаемые этим законом психометрические и диагностические вопросы всесторонне обсуждаются в заявлении Отделения оценки, измерения и статистики Американской психологической ассоциации, опубликованном в январском выпуске своего информационного бюллетеня *The Score*.

[2] Дополнительную информацию о паттернах развития маленьких детей с физическими недостатками и процедурные соображения по поводу их оценки можно найти в Wachs, & Sheehan (1988).

В одной из самых масштабных серий исследований, проведенных Службой тестирования в образовании, использовали стандартные и нестандартные версии *SAT* Совета колледжей и Общего теста *GRE*, предлагаемые четырем категориям поступающих в колледж и в аспирантуру: с нарушениями слуха, с нарушениями зрения, с трудностями в обучении (*learning disabled*) и с физическими нарушениями (Willingham et al., 1988). Исследуемые психометрические характеристики включали надежность, дифференцированное функционирование заданий, факторную структуру и другие показатели валидности, связанные с уровнями выполнения и прогнозирующей силой. Изучались также содержание тестов, временны́е параметры и приспособления к особенностям тестируемых. В общем, полученные результаты свидетельствуют о том, что процедурные приспособления сопоставимы со стандартной процедурой тестирования в большинстве отношений, включая значение показателей. Однако предсказание результатов учебной деятельности по показателям теста или по школьным отметкам оказалось менее точным для лиц с различными «неспособностями» (*disabilities*), чем для обычных абитуриентов. Кроме того, возник ряд сомнений по поводу факторной структуры и функционирования заданий некоторых адаптаций теста (R. E. Bennet, Rock, & Novatkoski, 1989; Rock, Bennet, & Jirele, 1988; Willingham, 1988). Вдобавок было обнаружено, что лимиты времени в нестандартных версиях являются относительно более снисходительными — результат, который подтверждает сомнительность практики засчитывания показателей как проходных при использовании этих версий. Таким образом разработка сопоставимых временны́х лимитов, основанных на эмпирических данных, стала насущной потребностью (см., например, Wainer, 1993 a, p. 9–10).

Остается еще много нерешенных психометрических и этических вопросов, касающихся тестирования лиц с различными дефектами. Признавая важность развертывания исследований в этой области, необходимо понять, что некоторые из этих проблем могут оказаться неразрешимыми вследствие их порождения *уникальностью* каждого индивидуума, обладающего неповторимой конфигурацией типов и уровней способностей, «неспособностей» и личных качеств. Тем не менее что касается практических целей, то уже сейчас можно отметить более высокий, чем когда-либо, уровень компетентности и чувствительности к потребностям лиц с различными видами дефектов, а также возросший уровень научно-методического обеспечения тестирования таких лиц. В добавление к этому, новые достижения в разработке аппаратуры, такие как создание аппаратов искусственной речи и других электронных устройств, управляемых компьютером, обеспечивают возможность использования разнообразных новаторских технологий тестирования, в том числе и тех, которые могли бы быть полезными в данной области (см., например, Educational Testing Service, 1992; Wilson, 1991).

В следующих разделах рассматриваются специальные вопросы тестирования лиц с тремя основными категориями физических недостатков, а именно с нарушениями слуха, зрения и двигательных функций.

Нарушения слуха.[1] Дети с ослабленным слухом (*hearing-impaired*) вследствие общего отставания в языковом развитии обычно отстают по показателям вербальных тестов, даже если вербальное содержание предъявляется визуально. Чем раньше у детей наступает глухота, тем сильнее это отставание. К счастью, современные дости-

---

[1] Анализ проблем и мнений, касающихся оценки детей с ослабленным слухом, можно найти в Bradley-Johnson, & Evans (1991), Y. Mullen (1992), Sullivan & Burley (1990).

жения в оценке слуховой деятельности сделали возможным точно диагностировать потерю слуха — и начинать реабилитационные процедуры — в течение первых месяцев жизни (Shah, & Boyden, 1991).

Некоторые из самых первых шкал действия создавались именно для тестирования глухих детей, например Шкала действия Пинтнера—Патерсона (*Pintner-Paterson Performance Scale*) и Шкала действия Артура (*Arthur Performance Scale*). В тестировании глухих детей часто используются специальные адаптации шкал Векслера. Большинство вербальных тестов можно применять при условии, что устные вопросы отпечатаны на карточках. Для сообщения инструкций в тестах действия были разработаны разнообразные методы (см., например, Sattler, 1988, 1992), и, фактически, самым широко используемым в США тестом интеллекта для детей с ослабленным слухом долгое время была Невербальная шкала *WISC-R.* И все же при введении подобных изменений в стандартные процедуры тестирования нельзя рассчитывать на то, что надежность, валидность и нормы теста останутся неизменными. Впрочем, благодаря широкому использованию шкал Векслера для обследования лиц с нарушениями слуха, здесь имеется обширная литература по психометрическим качествам этих шкал применительно к выборкам лиц с дефектами слуха (см., например, Braden, 1985; Maller, & Braden, 1993; Sullivan, & Schulte, 1992). В общем, эти исследования показывают, что имеет место существенное сходство в отношении как факторной структуры этих шкал, так и прогностической и конструктной валидности Невербальной шкалы для детей с ослабленным и нормальным слухом.

Все упоминавшиеся до сих пор тесты были стандартизованы на выборках испытуемых с нормальным слухом. Многие исследователи пришли к заключению, что когда уровни выполнения теста слабослышащими сопоставимы с таковыми у нормально слышащих, как в случае Невербальных шкал Векслера, то нет надобности в отдельных нормах для лиц с нарушениями слуха. В то же время нормы, полученные на глухих детях, бесспорно полезны в ряде ситуаций, имеющих отношение к их развитию в процессе обучения. Для удовлетворения этой потребности были предприняты отдельные попытки установить специальные нормы для существующих тестов, примером чего может служить стандартизация Невербальной шкалы для глухих детей *WISC-R* (*WISC-R Performance Scale for Deaf Children* — R. J. Anderson, & Sisco, 1977).

На более элементарном уровне был разработан и стандартизован на глухих и слабослышащих детях Тест способности к обучению Хискея—Небраска (*Hiskey-Nebraska Test of Learning Aptitude*). Этот тест требует индивидуального предъявления и рассчитан на детей от 3 до 17 лет. Фактор скорости выполнения был из теста исключен, поскольку было трудно объяснить смысл скорости маленьким глухим детям. Была также предпринята попытка охватить более широкое число интеллектуальных функций, чем это предусматривалось большинством тестов действия. В этом тесте для сообщения инструкций используют язык жестов и практические упражнения, а для установления раппорта — интересные, привлекающие детей задания. Все задания отбирались с учетом ограниченных возможностей глухих детей, причем окончательный выбор основывался главным образом на критерии возрастной дифференциации. Нормы устанавливались раздельно на выборках из 1079 глухих и 1074 слышащих детей. В руководстве к тесту, подробно рассматривающем процедуры, рекомендуемые при тестировании глухих детей, приведены параллельные инструкции для глухих и слышащих. Хотя нормы по тесту Хискея—Небраска явно устарели, сам тест имеет удовлетворительную надежность и валидность и до сих пор считается одним из лучших тестов для обследования детей с нарушениями слуха (Sullivan, & Burley, 1990).

На протяжении последних 50 лет рост знаний о последствиях глухоты для развития интеллекта был просто поразительным. Значительная часть этой истории освещена Брэденом (Braden, 1994) в исчерпывающем обзоре более 200 исследований глухих, в которых приняло участие свыше 170 000 человек. Брэден описывает многие интригующие результаты, полученные в этих исследованиях, включая и тот факт, глухие дети глухих родителей имеют показатели по тестам действия, превышающие нормы для детей с нормальным слухом. Хотя причины этого пока еще не полностью понятны, уже не остается сомнений в том, что глухота представляет собой гораздо более сложную переменную, чем считалось прежде. Этиология, степень, возраст наступления и обнаружения потери слуха, так же как и способ общения, уровень обучения (*educational placement*), состояние слуха родителей и наличие других дефектов — все эти факторы взаимодействуют и вносят свой вклад в различия познавательной деятельности у лиц с нарушениями слуха.

Нарушения зрения.[1] Тестирование слепых ставит перед исследователями совсем иные проблемы, чем те, с которыми они сталкиваются в работе с глухими. Устные тесты могут быть очень быстро адаптированы для слепых испытуемых, а вот применение тестов действия весьма затруднительно. В дополнение к обычному устному способу предъявления заданий могут быть использованы и другие тестовые методики, например магнитофонные записи. Кроме того, некоторые тесты, такие как Тест академической оценки (*SAT*) Совета колледжей, доступны в форматах с использованием крупного шрифта или шрифта Брайля. Последний метод несколько ограничен в своем применении из-за громоздкости тестовых материалов, напечатанных шрифтом Брайля, меньшей скорости чтения этого шрифта и из-за незнания шрифта Брайля некоторыми слепыми. Ответы тестируемых могут записываться либо с помощью шрифта Брайля, либо с использованием клавиатуры. Специально подготовленные ответы, выполненные выпуклым шрифтом на таблицах или карточках, вполне пригодны для использования в заданиях с множественным выбором, ответами типа «верно—неверно» и т. д. Разумеется, во многих индивидуально предъявляемых тестах испытуемые могут давать устные или жестикуляционные ответы.

Среди самых первых примеров тестов общего интеллекта, адаптированных для слепых, следует назвать тесты Бине. Первая редакция теста Хайеса—Бине для слепых создавалась на основе шкал Стэнфорд—Бине 1916 г. В 1942 г. была подготовлена Промежуточная форма теста Хайеса—Бине (*Interim Hayes-Binet*)[2] из варианта шкал Стэнфорд—Бине 1937г. (Hayes, 1942, 1943). Самая последняя адаптация — сопоставимая с Формой L-M Стэнфорд—Бине — Тесты интеллекта для слепых Перкинса—Бине (*Perkins-Binet Tests of Intelligence for the Blind*). Этот инструмент стандартизован на слабовидящих (*partially sighted*) и слепых детях и имеет отдельные формы для тестирования двух этих категорий детей (C. J. Davis, 1980).

Шкалы Векслера были также адаптированы для слепых испытуемых. Эти адаптации свелись, в сущности, к использованию вербальных тестов и отказу от тестов дей-

---

[1] Аналитические обзоры проблем и методов оценки детей с нарушениями зрения см. в Bradley-Johnson (1994), Fewell (1991), M. S. Moore, & McLaughlin (1992), Orlansky (1988).

[2] Разрабатываемый с самого начала как промежуточный вариант, поскольку для него была проведена лишь предварительная стандартизация, этот тест под таким названием и вошел в психологическую литературу.

ствия. Несколько заданий, не подходящих для слепых, были заменены приемлемыми вариантами. В общем, исследования детей с плохим зрением или с полной слепотой говорят о том, что эти состояния могут негативно сказываться на их когнитивном развитии, даже в вербальной области, из-за ограничений, которые они накладывают на широту и разнообразие детского опыта. Векслеровские профили детей с нарушениями зрения имеют сходный паттерн в разных исследованиях, и эти результаты свидетельствуют о том, что факторная структура задач у таких детей отличается от таковой у нормально видящих. Хотя показатели *IQ* не могут рассматриваться в качестве сколько-нибудь точных мер всей когнитивной деятельности детей с нарушениями зрения, в руках знающих и опытных специалистов шкалы Векслера они служат инструментом получения полезной диагностической информации о сильных и слабых сторонах функционирования интеллекта этих детей (Groenveld, & Jan, 1992).

Лишь очень немногие диагностические инструменты разрабатывались специально для оценки слепых и слабовидящих. Возможно, самым известным примером таких инструментов служит Тест способности слепых к обучению (*Blind Learning Aptitude Test* [*BLAT*] — Newland, 1979). *BLAT* — индивидуально проводимый тест, включающий адаптированные задания из теста Прогрессивные матрицы Равена и ряд других невербальных заданий, представленных в формате рельефных изображений. Акцент в данном тесте делается на процессе научения, а не на плодах прошлого обучения, которые могут создавать помехи слепому ребенку. Нормативные данные по *BLAT*, хотя и устаревшие, выгодно отличаются от нормативных данных, обычно доступных для такого рода инструментов. Сведения о надежности и валидности довольно скудные, и здесь требуются дополнительные исследования. Несмотря на это, *BLAT* может быть полезной составной частью (вместе с вербальными тестами) инструментария для оценки слепых детей младшего школьного возраста.

Более свежий пример — Тест интеллекта для детей с ослабленным зрением (*Intelligence Test for Visually Impaired Children* [*ITVIC*]) — еще находящийся в стадии доработки инструмент для комплексной оценки интеллекта слепых и слабовидящих детей, спроектированный группой исследователей из Нидерландов (Dekker, Drenth, Zaal, & Koole, 1990). *ITVIC* включает гаптические или тактильные варианты таких задач, как Складывание кубиков (*Block Design*), в состав батареи, в которую входят несколько невербальных и вербальных субтестов.[1] Этот тест требует дальнейших исследований на широкой выборке детей, однако уже предварительные данные позволяют рассчитывать на его валидность (Dekker, 1993; Dekker, Drenth, & Zaal, 1991; Dekker, & Koole, 1992).

Подобно всем другим обсуждаемым в этой главе состояниям, ослабленное зрение обнаруживает широкий диапазон градаций и весьма часто встречается в сочетании с другими дефектами. Поэтому принятие решения о том, использовать ли стандартные тесты, их адаптации или специально сконструированные тесты для слепых, зависит от целей оценки и уникальных особенностей обследуемого. В общем, пользователям тестов следует всегда помнить, что при таких модификациях тестов, как тактильное

---

[1] Гаптическая шкала интеллекта (*Haptic Intelligence Scale*) — аналогичный инструмент, разработанный для и нормированный на взрослых слепых в 1950-х — начале 1960-х гг. (Shurrager, & Shurrager, 1964). Эта шкала состоит из шести субтестов действия, построенных по образцу Шкалы интеллекта Векслера—Белльвью, а именно: Шифровка цифр (*Digit Symbol*), Сборка объекта (*Object Assembly*), Складывание кубиков (*Block Design*), Завершение объекта (*Object Completion*), Доска форм (*Pattern Board*) и Счет на предметах (*Bead Arithmetic*).

представление визуальных конструкций или увеличение лимитов времени, вряд ли можно рассчитывать на измерение тех же конструктов, что и при использовании оригинальных версий.

Нарушения моторики.[1] Лица с ортопедическими заболеваниями, способные нормально воспринимать слуховую и зрительную информацию, могут страдать такими тяжелыми расстройствами моторики, что для них оказываются недоступными ни устные, ни письменные ответы. Манипулирование с доской форм или другими материалами, используемыми в тестах действия, также может быть затруднено для них. Работа в условиях ограниченного времени или в незнакомом окружении часто усиливает имеющиеся у этих лиц нарушения моторики. А их повышенная утомляемость делает необходимым проведение тестирования короткими сериями.

Некоторые из наиболее тяжелых нарушений моторики свойственны страдающим церебральным параличом. Однако изучение этих случаев зачастую осуществлялось с помощью общих тестов интеллекта, таких как шкалы Стэнфорд—Бине. В таких исследованиях лица с наиболее тяжелыми формами ортопедических заболеваний обычно исключались как не поддающиеся тестированию, а в ходе тестирования часто допускались отступления от стандартной процедуры, с тем чтобы приспособить тест к возможностям реагирования обследуемого ребенка. Оба эти приспособления, разумеется, можно рассматривать лишь как временный выход из трудного положения.

Более удовлетворительный подход состоит в разработке инструментов тестирования, пригодных даже для лиц с самыми тяжелыми нарушениями моторики. В настоящее время для этой цели используют ряд специально созданных или адаптации существующих тестов, хотя данных об их нормативах и валидности по большей части недостаточно. Некоторые из обсуждающихся в следующем разделе тестов, первоначально предназначавшихся для использования в кросс-культурном тестировании, оказались пригодными и для обследования лиц с физическими недостатками. Были подготовлены адаптации Международной шкалы действия Лейтер (*Leiter International Performance Scale*) и Лабиринтов Портеуса (*Porteus Mazes*), пригодные для предъявления детям, страдающим церебральным параличом (Allen, & Collins, 1955; Arnold, 1951). В обоих адаптированных тестах тестирующий сам действует с тестовыми материалами, а тестируемый реагирует только определенными движениями головы. Прогрессивные матрицы Равена (ПМР) также служат пригодным для этой цели инструментом. Поскольку этот тест проводится без ограничений во времени и ответ может быть дан устно, письменно, указательным жестом или кивком, ПМР оказываются особенно подходящими для лиц с ортопедическими заболеваниями. Несмотря на гибкость и простоту способов ответа, тест ПМР включает задания широкого спектра трудности и обеспечивает довольно высокий верхний тестовый порог. В ряде работ сообщается об успешном использовании этого теста при изучении лиц с церебральным параличом и другими двигательными расстройствами (см., например, Capitani, Sala, & Marchitti, 1994).

Еще один тип тестов, допускающих в качестве ответа простые указательные жесты, представлен *словарными тестами в картинках (picture vocabulary tests)*. Эти тесты

[1] Обзор мер, полезных при оценке функций грубой моторики у маленьких детей, можно найти в H. G. Williams (1991), C. Robinson, & Fieber (1988) описывают процессуально ориентированный подход к оценке маленьких детей, использующий задачи Пиаже для сенсомоторного и дооперационального периодов.

обеспечивают быстрое измерение «пользования» словарным запасом, что делает их особенно пригодными для лиц, неспособных к отчетливому произношению слов (например, в случаях церебрального паралича) и для глухих. Поскольку они легки в применении и могут быть проведены примерно за 15 мин, словарные тесты в картинках можно также использовать как инструменты для экспресс-скрининга в ситуациях, где невозможно проведение полномасштабных индивидуальных тестов интеллекта.

Типичным образцом этого типа тестов является Словарный тест в картинках Пибоди. Его современная редакция (*PPVT-R* — Dunn, & Dunn, 1981) состоит из 175 листов иллюстраций, с четырьмя картинками на каждой. Предъявление каждой иллюстрации тестирующий сопровождает произношением вслух стимульного слова; тестируемый реагирует с помощью указательного жеста или каким-либо иным способом, выделяя на иллюстрации ту картинку, которая больше всего соответствует значению стимульного слова. Хотя полный тест охватывает возрастной диапазон от дошкольного детства до взрослости, каждому обследуемому предъявляют только те задания, которые соответствуют его уровню выполнения теста, определяемому по установленному соотношению серии успехов на одном и серии неудач на другом полюсе трудности. «Сырые» оценки могут переводиться в стандартные показатели ($M = 100$, $SD = 15$), процентили и станайны. Эти производные показатели наносят на карту с доверительными областями, покрывающими ± 1 *SEM* (стандартную ошибку измерения). Имеется возможность получения показателей в виде возрастных эквивалентов. Время проведения *PPVT-R* не лимитировано, но обычно на это требуется от 10 до 15 мин. Существуют две параллельные формы этого теста, использующие разные наборы изображений и стимульных слов.

PPVT-R был стандартизован на национальной выборке, включавшей 4200 детей и подростков в возрасте от 2,5 до 18 лет и 828 взрослых в возрасте от 19 до 40 лет. Психометрические характеристики теста являются вполне удовлетворительными (что касается соответствующих обзоров, см. McCallum, 1985; Wiig, 1985). Коэффициенты надежности, найденные путем оценки внутренней согласованности, сравнения взаимозаменяемых форм и повторного тестирования, колеблются от умеренных до высоких. Доказательства валидности *PPVT-R* опираются по большей части на прочную эмпирическую основу, заложенную в ходе исследований *PPVT*, с которым пересмотренная версия имеет среднюю корреляцию 0,70. Обзор свыше 300 исследований, использующих *PPVT*, обнаружил его высокие корреляции с другими словарными тестами, умеренные корреляции с тестами вербального интеллекта и академической способности, а также многообещающие связи с результатами по тестам учебных достижений. Корреляции имели сходную величину в различных популяциях, включая экономически неблагополучные группы населения и выборки лиц с разного рода «неспособностями» (*disabilities*) и психической задержкой. Показатели по *PPVT* отражают, отчасти, степень культурной ассимиляции респондента и степень воздействия на него нормативного американского английского языка.

Исследования с использованием самого *PPVT-R* показывают, что пересмотренная версия также имеет высокие корреляции с другими мерами вербального понимания (см., например, Elliott, 1990b, p. 235). Особенно интересное исследование *PPVT-R*, использующее структурную модель усвоения порядка слов, обеспечивает существенную поддержку конструктной валидности этого инструмента (L. T. Miller, & Lee, 1993). Третья редакция Словарного теста в картинках Пибоди — *PPVT-III* — вместе с тестом экспрессивной лексики (*expressive vocabulary test*), с которым он был конормирован, предположительно должна появиться в 1997 г.



Рис. 9–3. Проведение теста *CMMS* с ребенком

(Из Columbia Mental Maturity Scale: Guide for administering and interpreting. *Burgemeister et al., 1972, p. 11. Copyright © 1972 by Harcourt Brace Jovanovich, Inc. Воспроизводится с разрешения)*

Сходные процедуры проведения теста были внедрены в *тестах классификации изображений (pictorial classification tests),* что можно увидеть на примере Колумбийской шкалы умственной зрелости (*Columbia Mental Maturity Scale* [*CMMS*] — Burgemeister, Blum, & Lorge, 1972). Разработанная специально для использования с детьми, страдающими церебральным параличом, эта шкала включает 92 задания, каждое из которых содержит от 3 до 5 цветных рисунков, отпечатанных на большой карточке. От испытуемого требуется найти рисунок, который не принадлежит к классу других, обозначая свой выбор указательным жестом или кивком (см. рис. 9–3). Выборка стандартизации *CMMS* состояла из 2600 детей в возрасте от 3;6 до 9;11 и была репрезентативной относительно населения США по данным переписи 1960 г. Коэффициенты надежности эквивалентных половин и ретестовой надежности колеблются от 0,84 до 0,91. Корреляция со шкалой Стэнфорд—Бине, обнаруженная в группе из 52 дошкольников и первоклассников, равнялась 0,67. Корреляции с показателями теста достижений в выборках учащихся 1-х и 2-х классов попадают большей частью в интервал от 0,40 до 0,60. Как для более ранних, так и для современных форм *CMMS*, имеются обширные данные о валидности и применимости этого теста к разным группам инвалидов (см. *Tests in Print* II, III и IV). Однако вследствие устаревших норм и узкого диапазона оцениваемых способностей, применимость *CMMS* довольно ограничена.

Мультикультурное тестирование

Проблема. Тестированию людей, различающихся культурным происхождением, стало уделяться все большее внимание с начала 1950-х гг. Тесты необходимы для максимального использования людских ресурсов в развивающихся странах во многих частях мира. Быстро развивающейся системе образования в этих странах тестирование требуется как для организации приема в учебные заведения, так и для организации индивидуального консультирования. По мере развития промышленности появляется необходимость в тестах для отбора и распределения персонала, особенно в области обработки информации, машиностроения и автоматизации производства.

В Америке практические проблемы мультикультурного тестирования[1] связывались главным образом с культурами меньшинств, включенными в преобладающую культуру. В основном, интерес касался применимости имеющихся тестов к лицам, поставленным своей культурой в неблагоприятное положение. Следует ясно сознавать, что культурная ущербность (*cultural disadvantage*) — понятие относительное. Объективно между любыми двумя культурами или субкультурами существуют только различия. Каждая культура способствует развитию такого типа поведения, которое более приспособлено к ее ценностям и требованиям. Когда человек должен приспосабливаться и продвигаться в условиях культуры или субкультуры, отличающихся от той, в которой он воспитывался, то имеющиеся различия в культурах могут стать серьезным препятствием, а могут обернуться преимуществом.

Хотя интерес к кросс-культурному тестированию в значительной мере был вызван особенностями современного социального и политического развития, сама проблема была поставлена еще в 1910 г. Некоторые из первых кросс-культурных тестов создавались для тестирования эмигрантов, наплыв которых в США отмечался на рубеже двух столетий (Knox, 1914). Другие ранние формы тестов разрабатывались в рамках сравнительного изучения способностей людей, принадлежащих к относительно изолированным культурным группам. Эти культуры часто почти или совсем не соприкасались с западной цивилизацией, в рамках которой было разработано большинство психологических тестов.[2]

Традиционно, кросс-культурные тесты пытались исключить один или более параметров, по которым различаются культуры. Наиболее известным примером такого параметра служит *язык*. Если подлежащие тестированию культурные группы говорили на разных языках, то разрабатывались тесты, не требовавшие применения языка ни со стороны тестирующего, ни со стороны тестируемых. Если существенно варьировал уровень образования и преобладала неграмотность, исключались задания, требующие умения *читать*. Устная речь не исключалась из этих тестов, поскольку они предназначались для лиц, говорящих на общем языке. Другим параметром, по которому различаются культуры или субкультуры, является *скорость*. Не только темп ежедневной жизни, но мотивация и ценность быстрого выполнения заданий весьма заметно разнятся в разных национальных культурах, в этнических меньшинствах внутри одной нации и между городской и сельской субкультурами (см., например, Klineberg, 1928; R. R. Knapp, 1960; M. Womer, 1972). Соответственно в кросс-культурных тестах часто,

---

[1] Вместо термина «мультикультурное тестирование» широко употребляются такие термины, как «кросс-культурное тестирование» и «транскультуральное тестирование».

[2] Что касается примеров этих ранних тестов, см. Anastasi (1954, chap. 10).

хотя и не всегда, стремились элиминировать влияние скорости, увеличивая время выполнения заданий и не давая дополнительных баллов за более быстрое их выполнение.

Другие параметры, по которым различаются культуры, имеют отношение к *содержанию теста.* Так, например, материалом для неязыковых тестов и тестов для не умеющих читать служит информация, специфическая по отношению к конкретной культуре. Тесты могут требовать от испытуемого понимания назначений таких предметов, как скрипка, почтовая марка, ружье, перочинный нож, телефон, пианино или зеркало. Очевидно, лица, выросшие в относительно изолированных культурах, могут испытывать недостаток жизненного опыта для правильного ответа на такие задания. Главным образом для того, чтобы контролировать влияние параметров такого типа, и были разработаны первые классические «культурно-свободные» тесты. После краткого рассмотрения типичных тестов, предназначенных для устранения влияния одного или более перечисленных выше параметров, мы обратимся к анализу альтернативных подходов к кросс-культурному тестированию.

Типичные традиционные инструменты.[1] Пытаясь сконструировать тесты, пригодные для использования в различных культурах, психометристы использовали разнообразные процедуры, часть которых иллюстрируется рассматриваемыми в этом разделе тестами. Пересмотренная международная шкала действия Лейтер (Roid, & Miller, 1997) — индивидуально проводимый тест действия, впервые опубликованный в 1940 г. Шкала была подготовлена после применения в течение ряда лет в разных этнических группах на Гавайях. Впоследствии эта шкала была применена Портеусом к некоторым африканским группам и другими исследователями еще к нескольким национальных группам. Пересмотренная версия шкалы, выпущенная в 1948 г., основывалась на дополнительных результатах тестирования американских детей, учащихся средней школы и новобранцев времен Второй мировой войны. Редакция 1997 г. основана на выборках более 2000 типичных и нетипичных жителей США в возрасте от 2 до 20 лет. Отличительной чертой шкалы Лейтер, впоследствии заимствованной другими инструментами, является почти полное исключение речевых инструкций. Каждый тест начинается с самой легкой задачи того типа, с которым обследуемый сталкивается на протяжении всего этого теста. Понимание задач, которые даются индивидуально и без ограничения времени, рассматривается как часть теста. Весь графический стимульный материал предъявляется на специальных подставках, с соответствующим приспособлением для размещения карточек с ответами. Тестируемый отвечает на задачу, выбирая карточки с наиболее подходящими изображениями и помещая их на лоток для ответов, как можно увидеть на рис. 9–4.

Шкала Лейтер предназначалась для изучения широкого диапазона функций, аналогичных тем, для которых создавались вербальные шкалы. В ее современной форме этот диапазон существенно расширен, благодаря чему *LIPS-R* охватывает четыре области: Рассуждение (*Reasoning*), Визуализацию (*Visualization*), Внимание (*Attention*) и Память (*Memory*). К задачам, входящим на разных возрастных уровнях в области Рассуждения и Визуализации, относятся: рисуночные аналогии, завершение форм,

---

[1] Критический анализ некоторых невербальных средств измерения, обсуждаемых в этом и предыдущих разделах, так же как и других таких тестов, можно найти в Naglieri, & Prewett (1990).

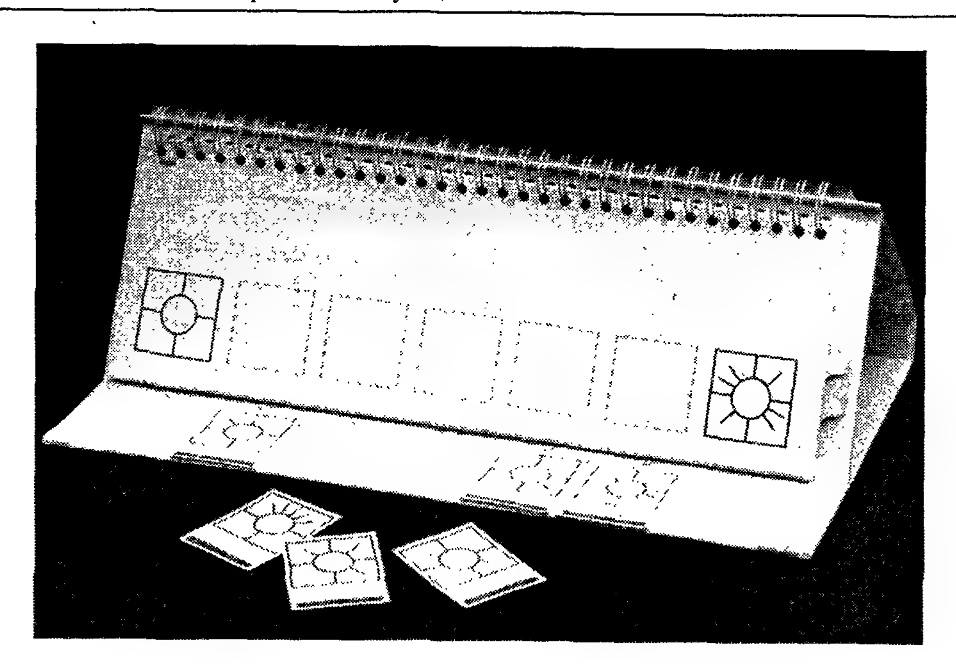


Рис. 9–4. Типичные материалы, используемые в Пересмотренной международной шкале действия Лейтер. Показанная здесь незавершенная задача из теста «Последовательный порядок» (*Sequential Order*) требует от испытуемого выбрать пять карточек из шести и разместить их в правильном порядке на лотке для ответов

(С любезного разрешения Stoelting Company)

установление сходства и последовательное упорядочивание (проиллюстрированное на рис. 9–4). Тесты областей Внимания и Памяти включают меры устойчивости и распределения внимания и разнообразные задачи на непосредственную и отсроченную память. Как и можно было ожидать, пересмотренная шкала Лейтер была существенно обновлена и стала более совершенной, чем ее ранние версии, в том, что касается психометрических характеристик. Например, градуировка уровней трудности в последней версии производилась на основе теории «задание – ответ» (*IRT*), а показатели *LIPS-R* уже не выражаются в виде традиционных коэффициентов *IQ*. В добавление к этому, наличие современных репрезентативных норм и расширенное содержание шкалы должны значительно повысить ее полезность. Новое руководство по *LISP-R* содержит сведения о различных типах надежности и данные о валидности.

Прогрессивные матрицы Равена (*Raven's Progressive Matrices* [*RPM*]) первоначально предназначались для измерения фактора *g* по Спирмену, или общего интеллекта (J. Raven, 1983; Raven, Raven, & Court, 1995). В соответствии с проведенным Спирменом теоретическим анализом фактора *g* этот тест требует главным образом выявления отношений между абстрактными элементами. Задания состоят из набора матриц, или композиций графических элементов, организованных в строки и столбцы, в каждой из которых один элемент пропущен. Задача состоит в том, чтобы выбрать подходящий элемент-вставку из заданного набора вариантов. Самые легкие задания требуют лишь точность различения, тогда как более трудные предполагают использование аналогий, перестановок, чередований паттерна и других логических отношений. Два образца типичных заданий из Стандартных прогрессивных матриц показаны на рис. 9–5. Тест

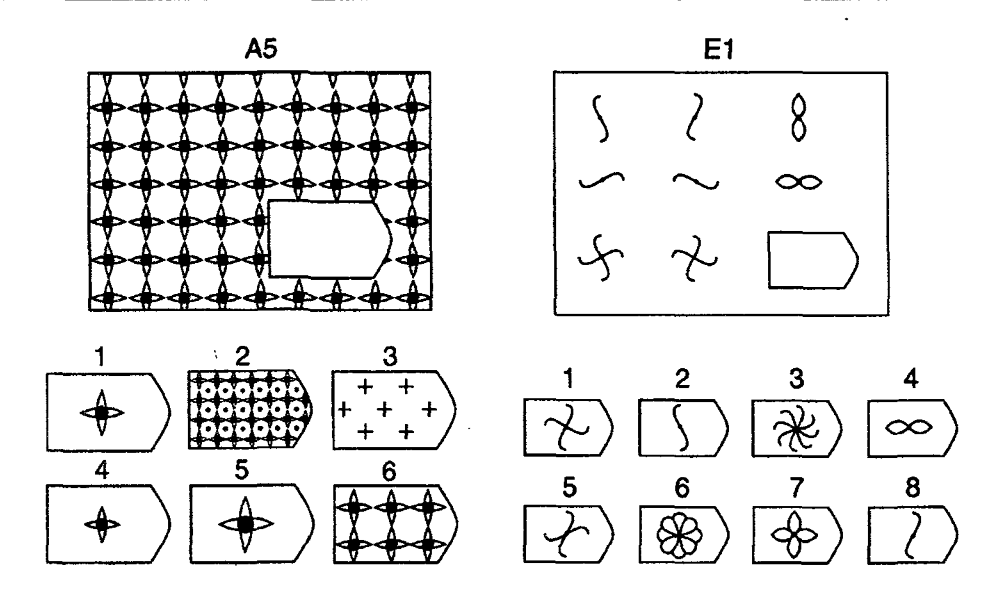


Рис. 9–5. Типичные задания из Стандартных прогрессивных матриц: одно легкое (A5) и одно трудное (E1)
(Воспроизводится с разрешения J. C. Raven Ltd.)

Равена обычно проводится без ограничений времени и может предъявляться индивидуально или группе испытуемых. Инструкции крайне просты и даются устно.

Имеется три формы Прогрессивных матриц Равена, различающихся по уровню трудности. Стандартные прогрессивные матрицы (*Standard Progressive Matrices* [*SPM* — 1996 Edition]) — форма, пригодная для обследования «средней» части человеческой популяции в возрастных границах от 6 до 80 лет. Более легкая форма — Цветные прогрессивные матрицы (*Coloured Progressive Matrices* [*CPM* — 1990 Edition]) — рассчитана на тестирование детей более младшего возраста и специфических групп, которые по разным причинам невозможно адекватно протестировать с помощью *SPM*. Нормы по *SPM* установлены для детей от 5,5 до 11,5 лет, а также для выборок лиц пожилого возраста без снижения интеллекта (от 60 до 89 лет) и умственно отсталых взрослых. Третья форма — Прогрессивные матрицы повышенной сложности (*Advanced Progressive Matrices* [*APM* — 1994 Edition]) — была специально разработана для тестирования подростков и взрослых, превосходящих средний уровень популяции.

Руководства для всех уровней Прогрессивных матриц Равена (*RPM*) выпускаются частями, которые можно приобрести по отдельности или в любой желаемой комбинации и в едином переплете. Часть 1 содержит общий обзор и обновлялась в последний раз в 1995 г.; обновление данных в других частях происходило в разные годы: от 1990 до 1996. Эти части содержат конкретные руководства для каждого из трех уровней *RPM*. В комплект тестов Равена входят также руководства по двум словарным тестам, стандартизованным для использования в сочетании с *RPM*. В последней части руководства приводятся сводные данные дополнительных исследований надежности и валидности, а также добавочные нормы, полученные в разных странах и на специфических популяциях (Court, & Raven, 1995). Пользователям доступны, кроме того, несколько дополнений с британскими данными стандартизации и нормативной инфор-

мацией, собранной в Северной Америке, Ирландии и Германии, а также аннотированная библиография более 2000 исследований с использованием *RPM*.[1]

Хотя к настоящему времени накопилось большое количество публикаций, посвященных результатам исследований *RPM*, эти исследования, вследствие преследуемых в них различных целей, крайне разобщены и разнородны. Авторы теста рекомендуют потенциальным пользователям выделять среди этого многообразия те исследования и те популяции, которые более всего отвечают их собственным интересам, но предупреждают, что все эти исследования существенно различаются по своей методологии, объемам выборок и качеству выполнения.

В общем, ретестовая надежность в группах старших детей и взрослых, умеренно однородных по возрасту, колеблется примерно от 0,70 до 0,90. Однако в области низких показателей надежность оказывается значительно меньше этих величин. Коэффициенты внутренней согласованности по больше части превосходят 0,80 и даже 0,90. Корреляции с вербальными и невербальными тестами интеллекта варьируют в пределах от 0,40 до 0,75, обнаруживая тенденцию быть выше с невербальными, чем с вербальными тестами. Исследования с умственно отсталыми и с различными профессиональными и образовательными группами свидетельствуют об удовлетворительной текущей валидности. Коэффициенты прогностической валидности относительно критериев успешности обучения оказываются несколько ниже соответствующих коэффициентов для обычных вербальных тестов интеллекта. Результаты факторного анализа, проведенного несколькими исследователями, говорят о том, что тест *RPM* имеет высокие нагрузки по общему фактору для большинства мер интеллекта (идентифицированному многими психологами как *g* Спирмена), но в то же время на выполнение этого теста влияют пространственная способность, индуктивное рассуждение, перцептивная точность и другие групповые факторы.

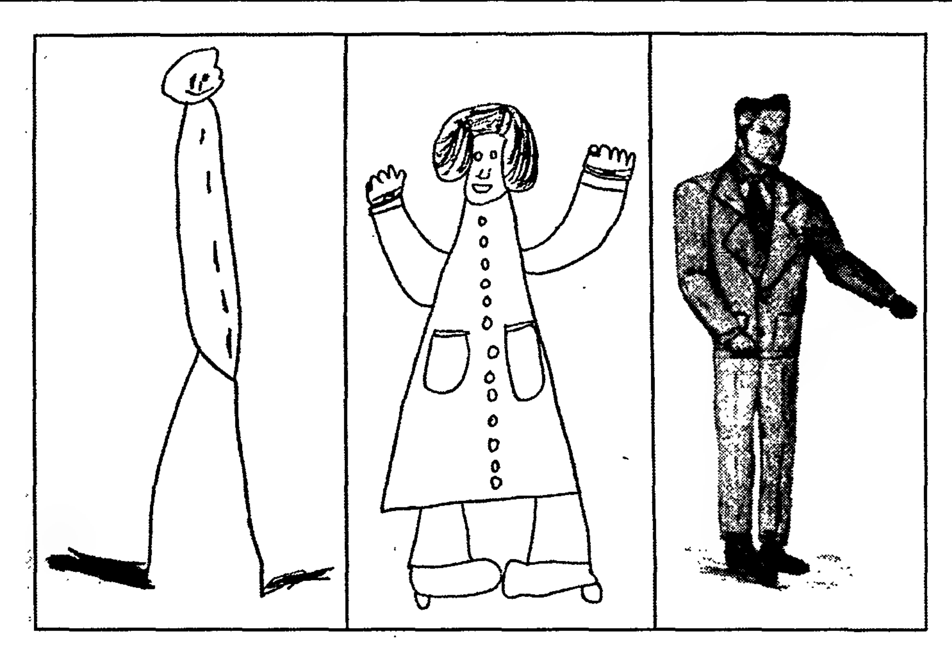
Иллюстрацией еще одного подхода к невербальному тестированию может служить тест Гудинаф «Нарисуй человека» (*Goodenough Draw-a-Man Test*), в котором испытуемому дают инструкцию «нарисовать мужчину и постараться сделать это как можно лучше». Этим тестом пользовались без изменений с момента его первоначальной стандартизации в 1926 г. до 1963 г. В 1963 г. его пересмотренная и расширенная версия была опубликована под названием Тест рисования Гудинаф—Харриса (*Goodenough-Harris Drawing Test* — D. B. Harris, 1963).

В нем, как и в исходном варианте, акцент делается на точности детской наблюдательности и на развитии понятийного мышления, а не на умении рисовать. При оценке учитывается, сколько и каких частей тела и деталей одежды изображает ребенок, как учтены пропорции, перспектива и другие особенности изображения. В итоге получилось 73 оцениваемых элемента, отобранных на основе возрастных различий, связи с суммарными показателями по этому тесту и с показателям группового теста интеллекта.

В пересмотренной версии шкалы тестируемых просили нарисовать женщину и самих себя. Подсчет баллов по шкале «Рисунок женщины» производится практически так же, как и по шкале «Рисунок мужчины». Шкала «Рисунок себя» разрабатывалась как проективный тест личности, но имеющиеся данные о ее применении нельзя на-

---

[1] Последнюю можно получить на диске или в виде распечатки у J. H. Court, по адресу, имеющемуся у издателей *RPM*.



Мужчина	Женщина	Мужчина:
Первичный показатель: 7	Первичный показатель: 31	Первичный показатель: 66
Хронологический возраст: 5;8	Хронологический возраст: 8;8	Хронологич. возраст: 12;11
Стандартный показатель: 73	Стандартный показатель: 103	Стандартн. показатель: 134

Рис. 9–6. Образцы рисунков, полученных в Тесте рисования Гудинаф—Харриса
(С любезного согласия Дейла Б. Харриса)

звать многообещающими.[1] Первичные показатели (в баллах) по каждой шкале преобразуются в стандартные показатели со средним $M = 100$ и $SD = 15$. На рис. 9–6 показаны три пояснительных рисунка, выполненных детьми в возрасте 5;8, 8;8 и 12;11, и соответствующие им первичные и стандартные показатели. Надежность Теста рисования Гудинаф—Харриса неоднократно исследовалась различными методами. Коэффициенты ретестовой надежности, надежности эквивалентных половин теста и надежности оценщика вполне удовлетворительны; влиянием обучения рисованию в школе на величину показателей, по-видимому, можно пренебречь (J. A. Dunn, 1967; D. B. Harris, 1963).

Помимо данных анализа заданий, собранных при разработке шкал, информацию о конструктной валидности дают корреляции с другими тестами интеллекта. Величина этих корреляций меняется в достаточно широких пределах, но большинство из них превышают 0,50. При исследовании детей, посещающих детский сад, оказалось, что Тест «Нарисуй человека» коррелировал выше с числовой способностью (*numerical aptitude*) и ниже со скоростью и точностью восприятия, чем это наблюдалось у учеников 4-го класса (D. B. Harris, 1963). Такие результаты говорят о том, что данный тест в разные возрастные периоды может измерять разные функции. Обе версии исполь-

[1] Другие проективные подходы к использованию рисунков человеческой фигуры обсуждаются в главе 15, включая методику Элизабет Коппиц (E. Koppitz), охватывающую как когнитивные, так и эмоциональные аспекты.

зовались в большом количестве исследований различных культурных и этнических групп, показавших, что выполнение этих тестов в большей степени зависит от различий в культурном происхождении, чем предполагалось первоначально. Деннис (Dennis, 1966), например, проанализировал сравнительные данные, полученные с помощью этого теста в 40 далеких друг от друга культурных группах, и обнаружил, что среднегрупповые показатели оказались весьма связанными со степенью представленности изобразительного искусства в каждой из культур.

Культурные различия в жизненном опыте обнаружились и в хорошо спланированном сравнительном исследовании мексиканских и американских детей с помощью теста Гудинаф—Харриса (Laosa, Swartz, & Diaz-Guerrero, 1974). В более позднем крупном исследовании этого теста в Иране средние показатели 6–13-летних детей получились несколько ниже американских норм, но при этом обнаружили хорошую возрастную дифференциацию и положительные корреляции с социоэкономическим статусом и мерами учебных достижений (Mehryar, Tashakkori, Yousefi, & Khajavi, 1987). Следует добавить, что такие результаты, полученные при использовании теста Гудинаф—Харриса, являются типичными результатами, получающимися при работе со всеми тестами, первоначально претендовавшими на роль «культурно-свободных» (*culture-free*) или «культурно-честных» (*culture-fair*) (Samuda, 1975, chap. 6).

Новая версия теста «Нарисуй человека», задуманная с целью обновления версии Гудинаф—Харриса и улучшения ее технических качеств, теперь доступна пользователям под названием «Нарисуй человека: Система количественной оценки» (*Draw A Person: A Quantitative Scoring System* [*DAP*] — Naglieri, 1988). *DAP* обеспечивает более свежие и детализированные нормы, но имеет несколько отличающуюся методику проведения и пересмотренную систему подсчета баллов, менее претенциозную по сравнению с системой Теста рисования Гудинаф—Харриса. Вдобавок ко всему, *DAP* включает нормативные данные, собранные на выборках чернокожих и испаноязычных детей. Однако несмотря на эти улучшения, данная версия была подвергнута критике за ее относительно узкий охват и отсутствие обоснования преимуществ новой системы подсчета баллов (Cosden, 1992).

В заключение вернемся к общей оценке обсуждаемых в этом разделе инструментов. Некоторые из них, хотя и разрабатывались первоначально для кросс-культурного тестирования, нашли основное применение в работе клинических и консультирующих психологов, — для получения информации, дополняющей данные, собранные с помощью таких инструментов, как шкалы Стэнфорд—Бине и Векслера, и для получения исходных данных в тестировании лиц с различными «неспособностями» (*disabilities*). Осознание этого факта привело к подготовке нового поколения таких средств измерения. Одно из них, доступное уже во второй редакции, — это Тест невербального интеллекта (*Test of Nonverbal Intelligence* [*TONI-2*] — L. Brown, Sherbenou, & Johnsen, 1990), который сходен по содержанию и диапазону применимости с *RPM* (что касается критических обзоров по *TONI-2*, см. K. R. Murphy, 1992 и Watson, 1992). В настоящее время проводится стандартизация других важных инструментов этого типа, которые предполагается выпустить в продажу в конце 1990-х гг.[1]

[1] Примером может служить Универсальный тест невербального интеллекта (Universal Nonverbal Intelligence Test) – авторы: B. A. Bracken & R. S. McCallum.

Подходы к кросс-культурному тестированию. Теоретически можно идентифицировать три подхода к разработке тестов для лиц, воспитанных в разных культурах или субкультурах, хотя на практике некоторые характерные особенности всех трех подходов могут сочетаться. Первый подход связан с подбором заданий, общих для множества различных культур, и валидизацией окончательного теста относительно локальных критериев в этих культурах. Это основной подход к созданию культурно-свободных тестов, хотя при его практической реализации вторичной валидизацией тестов в разных культурах часто либо просто пренебрегали, либо проводили ее неадекватно. Однако без этого этапа нельзя быть уверенным в том, что тест относительно свободен от элементов, свойственных определенной культуре. Более того, маловероятно, что вообще можно было бы разработать любой конкретный тест, полностью удовлетворяющий этим требованиям на широком спектре культур.

Тем не менее мультикультурные методы оценки необходимы для фундаментального исследования некоторых принципиальных вопросов. Один из таких вопросов касается универсальности психологических принципов и конструктов, полученных в рамках единственной культуры (Anastasi, 1958, chap. 18; Berry et al., 1992; Irvine, 1983; Irvine, & Carrol, 1980). Другой вопрос имеет отношение к роли средовых условий в формировании индивидуальных различий в поведении — проблема, которая может более эффективно изучаться в широком диапазоне средовой изменчивости, обеспечиваемой за счет выраженного различия культур. Исследования такого рода требуют инструментов, которые можно применять по крайней мере в частично сравнимых условиях различных культур. Меры предосторожности против неправильной интерпретации результатов, полученных с помощью таких инструментов, следует искать в подходящих для данной цели планах эксперимента и в основательном знакомстве исследователей с изучаемыми культурами или субкультурами. Что необходимо, так это установить специфические эмпирические переменные в любой данной культуре, которые могут быть связаны с социально значимыми различиями в поведенческом развитии, характеризующими такую культуру (J. W. Berry, 1983; Brislin, 1993; Segall, 1983; Whiting, 1976). Замечательный пример осуществления такой исследовательской программы — из области тестирования личности — можно найти в серии публикаций, посвященных кросс-культурному изучению тревожности (*Cross-Cultural Anxiety Series*). Этот цикл работ был посвящен исключительно измерению тревожности в разных культурах и оказался необычайно плодотворным в том, что касается расширения базы знаний об этом конструкте и о том, как тревога переживается людьми в разных частях света (см., например, Spielberger, & Diaz-Guerrero, 1990).

Второй подход состоит в том, чтобы создать тест внутри одной культуры и предъявить его людям с другими культурными корнями. В этом случае мы должны избегать рассматривать любой тест, разработанный в рамках одной культуры, как универсальную мерку для измерения «интеллекта» или других конструктов. Не следует также предполагать, что низкий показатель по такому тесту имеет одинаковое причинное объяснение для двух лиц, принадлежащих к разным культурам. Что мы действительно *можем* установить с помощью такого подхода, так это культурную дистанцию между группами, а еще степень аккультурации индивидуума и его готовность к получению образования и профессиональной деятельности, специфичных для данной культуры. Некоторые исследователи пытались придать особое значение тому, что культурная среда, в которой воспитывается человек, влияет на приобретаемые им когнитивные навыки и знания. Ранние примеры включают тест распознавания следов (*footprint*

recognition test), стандартизованный на австралийских аборигенах (Porteus, 1931), и Тест «Нарисуй лошадь» (*Draw-a-Horse Test*), стандартизованный на детях индейцев пуэбло (DuBois, 1939).

Согласно третьему подходу, внутри каждой культуры могут разрабатываться специфические тесты (или основательные адаптации существующих тестов), которые должны валидизироваться относительно локальных критериев и использоваться только в соответствующей культуре. Иллюстрацией этого подхода служит разработка тестов для отбора военного и промышленного персонала в определенных культурах. Конкретный пример дает программа по разработке тестов, реализуемая в некоторых развивающихся странах Азии, Африки и Латинской Америки при поддержке Агентства международного развития (Schwarz & Krug, 1972). В таких случаях тесты валидизируются относительно конкретных образовательных и профессиональных критериев, для прогнозирования которых эти тесты создаются, а их выполнение оценивается исходя из локальных норм. Каждый тест применяется только в той культуре, где он был разработан, и не используется для кросс-культурных сравнений. Однако если предсказываемые критерии имеют отношение к технологии, вероятно, востребованным окажется «интеллект западного типа», — и тесты будут отражать направление, в котором развивается конкретная культура, а не свойственные ей в настоящее время особенности. Вдобавок ко всему, как показывает недавний обзор использования тестов в мире, современная действительность такова, что в целом чаще всего применяются, — по крайней мере, при обследовании детей и молодежи, — тесты, сконструированные в США и Европе. Фактически, среди всех охваченных этим обзором государств, наименее развитые страны, которые, вероятно, в наибольшей степени отличаются от Соединенных Штатов и европейских стран, более других опираются на зарубежную технологию тестирования (Hu, & Oakland, 1991; Oakland, & Hu, 1992).[1]

К настоящему времени накопилась обширная литература по психологическому тестированию культурных меньшинств внутри плюралистических обществ, таких как США, Израиль и Нидерланды (см., например, Bleichrodt, & Drenth, 1991; Duran, 1989; Figueroa, 1990; Hessel, & Hamers, 1993; Samuda, Kong, Cummins, Lewis, & Pascual-Leone, 1991; Zeidner, 1988). В данной книге мы обращаемся к этому материалу всякий раз, когда его можно ясно и сжато изложить. Так, в главе 18 центром рассмотрения станут вопросы социальной и этической ответственности и соблюдения интересов тестируемых при применении тестов в работе с культурными меньшинствами. Технические психометрические проблемы систематической ошибки тестов и взаимодействия «задание х группа» рассматривались в главах 6 и 7. А в этой главе акцент был сделан на инструментах, разрабатываемых для кросс-культурного тестирования способностей. Проблемы в интерпретации результатов кросс-культурного тестирования, вместе с современными тенденциями, будут рассмотрены в главе 12.

В наши дни мультикультурное тестирование постепенно уходит от конструирования специальных тестов и все больше сосредоточивается на роли тестирующего в процессе проведения обследования. По существу, в обязанности тестирующего входит: 1) получение информации о культурном происхождении тестируемого; 2) выбор тес-

---

[1] Учитывая существующее положение дел, Международная комиссия по тестам (*International Test Commission*) подготовила тщательно продуманный и ясный набор методических рекомендаций по адаптации образовательных и психологических тестов (Hambleton, 1994, 1996; Van de Vijver, & Hambleton, 1996). Многие из этих вопросов рассмотрены в статье Geisinger (1994).

та, наиболее пригодного для той цели, ради которой он используется; 3) эффективное проведение теста с конкретным испытуемым; 4) интерпретация результатов теста с учетом истории жизни испытуемого и того контекста (профессионального, образовательного, общественного и т. д.), в котором оцениваются его квалификационные данные. Эти функции роли тестирующего будут дополнительно обсуждаться в главе 12.

Оценка среды. Хотя изучение традиционных кросс-культурных тестов представляет исторический интерес и, в связи с этим, улучшает понимание происхождения и природы современных тестов, быстро растущие контакты между мировыми культурами радикально меняют потребность в таких тестах. Все больше и больше эффективных тестов будет разрабатываться (или адаптироваться) в конкретных культурах и для совершенно конкретных целей — например, для применения в сферах образования, трудоустройства или консультирования. Бесперспективность поисков универсального теста человеческого интеллекта стала очевидной вследствие растущего понимания значительного вклада в его формирование условий и истории жизни конкретного человека. А это привело к росту активности в области оценивания среды функционирования индивидуума.[1]

Традиционный подход к оценке среды человека опирался на довольно общий, комплексный индекс социоэкономического уровня. Социологи пользовались сложными методиками определения принадлежности индивидуума к социальному классу (Warner, Meeker, & Eells, 1949). Однако проще и быстрее вычисляемые индексы оказались равно эффективными, давая результаты, весьма близкие к получаемым с помощью трудоемких социологических методов. В действительности, достаточно близкую аппроксимацию социоэкономического уровня можно получить на основе учета профессии основного кормильца в семье. Было сконструировано несколько грубых шкал для классификации родительских профессий по уровням; в некоторых из них информация о профессии объединяется с уровнем образования родителей, как в широко используемом двухфакторном Индексе социального положения (*Two-Factor Index of Social Position*). Этот индекс, впервые описанный Холлиншедом (Hollingshead, 1957), можно найти в разных источниках (например, Bonjean, Hill, & McLemore, 1967; Hopkins, & Stanley, 1981). Были разработаны и более объективные методы регистрации сведений о профессиональной деятельности и выведения на их основе индекса профессионального уровня (Duncan, 1961; Stricker, 1985).

Главное ограничение традиционных глобальных индексов проистекает из того, что они классифицируют среды в одномерном континууме: лучше — хуже или выше — ниже. На самом деле среды различаются по подкрепляемому ими конкретному поведению и, следовательно, по их воздействию на специфические индивидуальные характеристики (см., например, McAndrew, 1993). Поэтому оптимальные среды для развития атлетических навыков, школьных умений, креативности и социальной конформности могут принципиально различаться. Ценное руководство по эмпирическому подходу к классификации и описанию условий внешней среды, влияющих на поведение человека, можно найти в новой редакции пионерской работы Роджера Баркера по экологической психологии (Schoggen, 1989).

Кросс-культурное тестирование выдвигает на первый план важную роль, которую родительское поведение и домашняя обстановка играют в интеллектуальном разви-

[1] Этот вопрос обсуждается более подробно в главе 12.

тии растущего ребенка (см., например, М. Н. Bornstein, 1991). Сейчас также признается, что такие средовые различия не ограничиваются ясно определимыми культурными или этническими популяциями, но могут оказывать существенное влияние на психологическое развитие любого человека. Кроме того, изучаемые среды требуют более конкретного определения на основе поощряемого ими специфического поведения. Более точной оценке психологического влияния различных домашних условий и семейной атмосферы было уделено повышенное внимание.

В наше время пользователям доступно довольно много мер и разного типа методик оценки семьи и домашних условий (Bradley, & Brisby, 1993; Paget, 1991). Хорошо известный и широко используемый инвентарь домашней среды называется «Обследование семьи для оценки условий жизни» (*Home Observation for Measurement of the Environment* [*HOME*] — В. М. Caldwell & Bradley, 1984). Этот инструмент нацелен на выявление типов стимуляции и родительского поведения в домашней обстановке, которые способствуют когнитивному развитию (Bradley Caldwell, 1984; В. М. Caldwell, & Bradley, 1978; J. H. Stevens, & Bakeman, 1985). Инвентарь *HOME* в настоящее время доступен в трех версиях, предназначенных для обследования семей с детьми трех возрастных категорий: от рождения до 3 лет, от 3 до 6 лет и от 6 до 10 лет. *HOME* позволяет получить показатели по нескольким шкалам, оценивающим такие переменные, как обеспечение ребенка подходящим игровым материалом, разнообразие стимуляции, языковая стимуляция, поощрение социальной зрелости и учебного поведения (что касается обзора, см. Boehm, 1985). Индексы социоэкономического статуса (*SES*) семей младенцев коррелируют с интеллектуальной деятельностью в раннем детстве также или даже сильнее, чем показатели *HOME*. Однако сочетание *SES* и показателей HOME может повышать предсказуемость интеллекта при определенных обстоятельствах (см., например, D. L. Johnson et al., 1993). К тому же переменные, оцениваемые с помощью инвентаря *HOME* и других сходных инструментов, могут добавить уникальную и ценную информацию к оценке детей, производимой для многих других целей.

10 ГРУППОВОЕ ТЕСТИРОВАНИЕ

В то время как индивидуальные тесты, такие как шкалы Стэнфорд—Бине и Векслера, находят свое основное применение в клинике, групповые тесты используются преимущественно в системе образования, гражданских службах, в промышленности и армии. Напомним, что массовое тестирование началось в США во время Первой мировой войны с разработки армейских тестов альфа и бета. Армейский альфа представлял собой вербальный тест, предназначенный для общего отбора и распределения новобранцев. Армейский бета был неязыковым тестом и предназначался для не владеющих английским или неграмотных новобранцев, которых невозможно было протестировать с помощью формы альфа. Эти тесты явились своего рода образцом для последующего развития большого числа групповых тестов для гражданского населения.

Пересмотренные гражданские формы обоих армейских тестов продолжали использоваться еще не один десяток лет после окончания войны. В армии США позже был разработан Квалификационный тест вооруженных сил (*Armed Forces Qualification Test* [*AFQT*]) в качестве средства предварительного отбора, с последующим использованием комплексных классификационных батарей способностей для распределения военнослужащих по соответствующим армейским специальностям. AFQT обеспечивает единый показатель, получаемый на основе выполнения равного количества заданий на выявление словарного запаса, арифметических и механических способностей, понимания пространственных отношений. Еще позднее была разработана Батарея профессиональной пригодности Вооруженных сил (*Armed Services Vocational Aptitude Battery* [*ASVAB*]) для использования во всех родах войск в качестве комбинированного инструмента отбора и классификации военнослужащих. Некоторые субтесты *ASVAB* служат для оценки общей пригодности к воинской службе. Что касается распределения персонала, то каждая армейская служба выбирает и комбинирует субтесты таким образом, чтобы они в наибольшей степени отражали требования конкретной воинской специальности.

В этой главе мы сначала рассмотрим принципиальные различия между групповыми и индивидуальными тестами. За этим последует беглый обзор начинающих появляться процедур индивидуально приспособленного тестирования в группах и использования компьютеров в программах тестирования. Затем мы приведем несколько свежих примеров групповых тестов широкого назначения. В заключение мы рассмотрим

главную современную тенденцию в разработке и применении тестов, которая отчетливо проявляется как в области групповых, так и в области индивидуальных тестов, обсуждавшихся в главе 8. Эта тенденция — к слиянию тестов, первоначально разрабатываемых в качестве общих мер единственной широкой способности (например, интеллекта или способности к обучению), с комплексными батареями способностей. Все больше тестов способностей адаптируется в целях обеспечения гибкости использования, в результате чего один измерительный инструмент может давать показатели разного уровня обобщенности — от общих до специфических, отвечая широкому разнообразию целей и ситуаций тестирования.

Групповые тесты в сравнении с индивидуальными

Типичные различия в конструкции тестов. Групповые тесты неизбежно отличаются от индивидуальных формой и организацией заданий. Хотя в них и можно было бы применять вопросы, допускающие неограниченное количество ответов в свободной форме, — как это имело место в первых групповых тестах, — в типичных современных групповых тестах используются *задания с множественным выбором (multiple-choice items)*. Это изменение очевидно было вызвано требованиями единообразия и объективности при подсчете баллов. Другое важное различие между традиционными индивидуальными и групповыми тестами состоит в контроле трудности заданий. В индивидуально проводимых тестах тестирующий следует правилам определения начального, базального и предельного уровней, чтобы обеспечить каждому тестируемому проверку с помощью заданий, соответствующих его уровню способности. В групповых тестах сходные по содержанию задания располагаются в порядке возрастающей трудности в виде относительно самостоятельных, *разделенных во времени субтестов (separately timed subtests)*. Такая организация заданий дает тестируемому возможность попробовать свои силы в каждом их типе (например, на словарный запас, арифметику, пространственные отношения и т. д.) и выполнить более легкие из них до того, как приступить к более трудным, на попытки справиться с которыми у него, в противном случае, могла бы уйти впустую значительная часть отведенного времени.

Однако практическая трудность, встречающаяся при использовании раздельных субтестов, состоит в том, что менее опытные и менее внимательные пользователи могут допускать ошибки временно́й организации тестирования (*timing errors*). Такие ошибки, по-видимому, чаще встречаются и имеют более серьезные последствия при установлении нескольких коротких лимитов времени (для каждого субтеста), чем при работе с одним, достаточно большим временны́м лимитом (для теста в целом). Чтобы совместить использование одного лимита времени на весь тест с таким расположением заданий, которое позволило бы всем тестируемым испробовать все типы заданий на последовательно возрастающих уровнях трудности, в некоторых тестах применяется *спиральное расположение заданий (spiral-omnibus format)*. Одним из первых примеров такого расположения заданий дают Самоприменяемые тесты умственных способностей Отиса (*Otis Self-Administering Tests of Mental Ability*), в которых, как указывает их название, предпринята попытка свести роль проводящего обследование к минимуму. В тесте со спиральным расположением заданий самые легкие задания каждого типа предъявляются первыми, затем идет следующий по степени трудности ряд заданий каждого типа и т. д., примерно так, как это показано ниже:

Ответ

1. Противоположным ненависти является: а) вражда, б) страх, в) любовь, г) дружба, д) радость .... ( )
2. Если 3 карандаша стоят 25 центов, сколько карандашей можно купить на 75 центов? .... ( )
3. У птицы не всегда бывают: а) крылья, б) глаза, в) ноги, г) гнездо, д) клюв .... ( )
4. Противоположным чести является: а) слава, б) бесчестье, в) трусость, г) страх, д) поражение .... ( )

Для того чтобы избежать необходимости повторять инструкции для каждого задания и сократить число переключений с одной установки на другую, требуемых от испытуемого инструкциями к заданиям разных типов, в некоторых тестах по спирали располагаются не единичные задания, а блоки из 5–10 заданий.

Преимущества группового тестирования. Групповые тесты разрабатываются в первую очередь как инструменты массового тестирования. По сравнению с индивидуальными тестами у них есть свои достоинства и свои недостатки. Позитивной стороной групповых тестов является возможность проводить их одновременно с таким большим количеством людей, которое только можно удобно разместить в пригодном помещении, размеры которого ограничиваются, пожалуй, лишь пределом слышимости голоса тестирующего, пользующегося микрофоном. Именно развитие методов группового тестирования сделало возможным реализацию программ массового тестирования. Благодаря использованию заданий теста в отпечатанном виде и простых ответов, легко фиксируемых в тестовой тетради, на бланке ответов или с помощью компьютера, отпала необходимость взаимодействия тестирующего и тестируемого один на один.

Еще одной особенностью группового тестирования, облегчившей проведение массовых обследований, явилось значительное упрощение функций проводящего тест. В отличие от всесторонней подготовки и большого опыта, необходимых пользователю, например, при тестировании по шкале Стэнфорд–Бине, для предъявления большинства групповых тестов от него требуется лишь умение зачитывать простые инструкции испытуемым и точно соблюдать время. Конечно, желательно проводить с пользователями групповых тестов предварительные тренировочные занятия, так как неопытность может стать причиной отклонения от стандартизованной процедуры тестирования и тем самым сказаться на результатах теста. В то же время при групповом тестировании могут быть обеспечены более единообразные условия, чем при индивидуальном, поскольку роль тестирующего сведена к минимуму. Использование магнитофонных записей инструкций и компьютерного предъявления заданий теста открывает дополнительные возможности для процедуры стандартизации и устранения фактора различий между проводящими массовое тестирование специалистами. Подсчет показателей при групповом тестировании обычно носит более объективный характер и может быть выполнен даже вспомогательным персоналом. В настоящее время большинство групповых тестов вообще предполагает компьютерную обработку результатов.

Кроме того, групповые тесты, как правило, позволяют получить более точные и надежные нормы, чем индивидуальные. Вследствие относительной легкости и быст-

роты сбора данных с помощью групповых тестов, обычно в процессе их стандартизации тестированию подвергаются большие, репрезентативные выборки. Для большинства современных стандартизованных групповых тестов нет ничего необычного в том, что их нормативные выборки насчитывают от 100 000 до 200 000 человек, в отличие от 1000 (максимум — 8000) случаев, с трудом накопленных в ходе стандартизации даже наиболее тщательно разработанных индивидуальных шкал интеллекта.

Недостатки группового тестирования. Хотя групповые тесты обладают некоторыми желательными свойствами и выполняют практически незаменимую функцию в современном тестировании, следует отметить и их ограничения. При групповом тестировании у проводящего тест гораздо меньше возможностей для того, чтобы установить раппорт с испытуемыми, добиться от них сотрудничества и поддерживать их интерес. Любые временные состояния испытуемого, такие как нездоровье, утомление, беспокойство или тревога, которые могут помешать выполнению заданий, гораздо труднее обнаружить при групповом тестировании, чем при индивидуальном. В целом лица, непривыкшие к тестированию, скорее покажут более низкие результаты в групповых тестах, нежели в индивидуальных. Существуют данные, свидетельствующие о том, что дети с нарушениями эмоциональной сферы лучше выполняют индивидуальные тесты, чем групповые (Bower, 1969; Willis, 1970).

С другой стороны, групповые тесты неоднократно подвергались нападкам за ограничения, налагаемые на ответы испытуемых. Особенно критикуются задания с множественным выбором ответов и такие стандартные типы заданий, как аналогии, нахождение сходства и классификация (Hoffman, 1962; LaFave, 1966). Ряд критических замечаний носит оригинальный характер и стимулирует совершенствование заданий групповых тестов. Одно из направлений этой полемики касается того, что такие задания ставят в невыгодное положение тех, кто блестяще и оригинально мыслит, кто ищет и стремится выразить в ответах необычный смысл. Заметим, кстати, что если это и происходит, то очень редко, о чем говорят анализ заданий и данные по валидности. Если все же такое случится в одном или двух заданиях предъявляемого индивидууму теста, то едва ли окажет заметное влияние на совокупный показатель данного испытуемого. Некоторые критики, что характерно для подхода Пиаже (Sigel, 1963), указывают на важность анализа ошибок и выяснения причин, которые побуждают индивидуума выбирать определенный ответ. Несомненно, групповые тесты почти или совсем не позволяют непосредственно наблюдать поведение испытуемых и устанавливать источник нетипичного выполнения тестов. По этой и другим причинам, когда принимаемое по результатам тестирования решение важно для испытуемого, желательно дополнить результаты группового тестирования либо индивидуальной проверкой неясных случаев, либо информацией, полученной из других источников.

Еще одним ограничением традиционного группового тестирования является его недостаточная гибкость, поскольку каждый обследуемый тестируется одинаково по всем заданиям, хотя отводимое для тестирования время может быть использовано более эффективно, если каждый испытуемый сосредоточит свои силы на заданиях, соответствующих его уровню способностей. Более того, такая процедура могла бы помочь избежать скуки при выполнении слишком легких заданий, с одной стороны, а с другой — способствовала бы снятию фрустрации и тревожности при попытке выполнить задания, превышающие по сложности уровень способностей индивидуума. Индивидуальные тесты в типичных случаях позволяют тестирующему выбирать за-

дания на основе предшествующих ответов тестируемого. Это различие между индивидуальными и групповыми тестами особенно важно, когда тест предназначен для охвата широкого диапазона измеряемой способности.

Адаптивное тестирование и компьютеризованное проведение тестов

Адаптивное тестирование. Индивидуально адаптируемые тесты. С тем чтобы объединить некоторые достоинства индивидуального тестирования с преимуществами группового, опробуется ряд методик. Основной интерес до сих пор сосредоточивался на способах приспособления набора заданий к характеристикам ответов отдельных испытуемых. Во все увеличивающейся литературе, посвященной этой проблеме, такой подход назывался по-разному: адаптивное, последовательное, разветвленное, специализированное, индивидуализированное, программируемое, динамическое или зависящее от ответа тестирование. Хотя вполне можно создавать тесты типа «карандаш—бумага», включающие такие адаптивные процедуры (Cleary, Linn, & Rock, 1968; Lord, 1971), сами эти методики идеально подходят для компьютеризованного проведения тестов.

Адаптивное тестирование может строиться на основе широкого множества процедурных моделей (DeWitt, & Weiss, 1974; Larkin, & Weiss, 1974; Weiss, 1974; Weiss, & Betz, 1973). Простой пример тестирования в две стадии приведен на рис. 10–1. В этом гипотетическом тесте все испытуемые проходят тест, состоящий из 10 заданий самой разной степени трудности, с целью определения маршрута дальнейшего обследования. В зависимости от успешности выполнения этого теста-маршрутизатора испытуемому предъявляется один из трех различных по трудности измерительных тестов, каждый из которых состоит из 20 заданий. Таким образом, испытуемый выполняет только 30 заданий, в то время как тест в целом содержит 70 заданий.

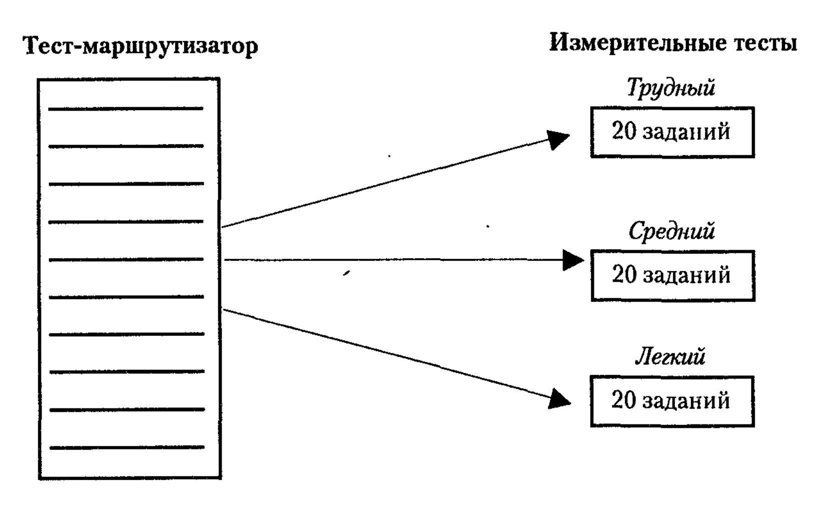


Рис. 10–1. Двустадийное адаптивное тестирование с тремя уровнями измерения. Каждый испытуемый проходит тест-маршрутизатор и один из трех измерительных тестов

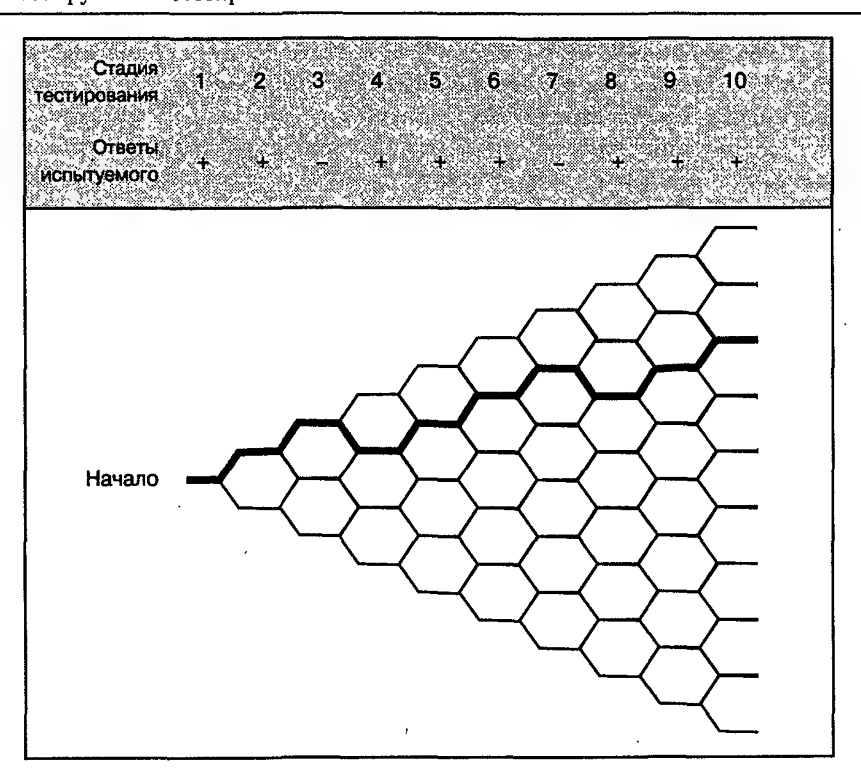


Рис. 10–2. Пирамидальная модель тестирования. Жирной линией показан маршрут обследования одного испытуемого, чьи результаты выполнения заданий приведены в верхней части рисунка

Иная организация заданий иллюстрируется пирамидальным тестом, изображенным на рис. 10–2. В этом случае все испытуемые начинают с задания средней трудности. Если ответ испытуемого на это задание правилен, то ему предъявляется следующее по степени трудности задание; если неправилен, то следующее по степени легкости. Процедура повторяется после каждого ответа испытуемого до тех пор, пока он не даст 10 ответов. Это пример 10-стадийного теста, в котором каждому испытуемому предъявляется 10 из 55 входящих в тест заданий. Жирная линия на рис. 10–2 показывает маршрут обследования конкретного испытуемого, ответы которого на предлагаемые задания отмечены вверху знаками + (правильно) и — (неправильно).

Компьютеризованное адаптивное тестирование (КАТ). Некоторые варианты обеих моделей адаптивного тестирования, примеры которых показаны на рис. 10–1 и 10–2, были реализованы как в форме «карандаш—бумага», так и на базе компьютера. Однако более сложные модели, не предусматривающие заранее установленного, фиксированного порядка предъявления заданий, допускают реализацию только в форме компьютеризованного адаптивного тестирования (Embretson, 1992; B. F. Green, 1983; Wainer et al., 1990). В основу этих процедур КАТ положены методы описанной в главе 7 теории «задание — ответ» (*IRT*), которые используются для составления комплекта заданий, проведения тестирования конкретных испытуемых и подсчета индивидуальных показателей. Для каждого задания теста существует оценка способности, тре-

буемой для его выполнения с вероятностью 0,50. Эта оценка способности и служит тем показателем, который индивидуум получает за правильное выполнение данного задания. Такой показатель отражает уровень трудности, различительную способность и вероятность угадывания правильного ответа для данного задания. Кроме того, для каждого задания имеется информационная функция, показывающая точность измерения. Информационная функция теста, представляющая собой сумму информационных функций заданий, выполняет ту же роль, что и традиционная стандартная ошибка измерения (*SEM*). После ответа испытуемого на каждое задание компьютер выбирает для него следующее задание с учетом всей «предыстории» его ответов. Добавление новых заданий в процессе тестирования продолжается до тех пор, пока информационная функция теста не достигает заранее установленного стандарта. Таким образом, при обследовании всех испытуемых достигается одинаковый уровень точности измерений.

Показатель конкретного испытуемого основывается не только на количестве правильно выполненных заданий, но отражает уровень трудности и другие психометрические характеристики этих заданий. Совокупный тестовый показатель выводится на основе оценок способности, соответствующих каждому выполненному заданию. Эта оценка способности исправляется и уточняется с добавлением каждого нового задания до тех пор, пока не достигается заданная точность измерения. Такие показатели будут сопоставимы у всех лиц, обследованных с помощью комплекта входящих в тест заданий, независимо от специфического набора заданий, предъявленных каждому испытуемому. Существующие на сегодняшний день процедуры конструирования инструмента КАТ можно существенно облегчить за счет использования ряда доступных компьютерных программ, таких как *MicroCAT*, распространяемых *ASC* (*Assessment Systems Corporation*).[1]

В общем, исследования, проведенные разными методами, показывают, что индивидуализированное адаптивное тестирование может давать столь же надежные и валидные результаты, как и общепринятые тесты, однако при существенно меньшем числе заданий и значительной экономии времени. Кроме того, оно обеспечивает большую точность измерения для испытуемых, находящихся на верхнем и нижнем краях диапазона способности, охватываемого тестом (Lord, 1970; 1971a; 1971b; 1971c; Weiss, 1982). Было также проведено важное исследование, показавшее, что корреляции между правильно сконструированными КАТ формами тестов и их бланковыми формами (типа «карандаш — бумага») почти столь же высоки, как коэффициенты надежности большинства тестов. Такие результаты говорят о том, что одни и те же конструкты по существу можно измерять с помощью обеих форм тестов (Mead, & Drasgow, 1993). В то же время есть ситуации тестирования, для которых КАТ не подходит, например когда используются тесты скорости и скрининг-тесты, распределяющие испытуемых по группам на основе критического показателя (Wainer, 1993b). Особое внимание уделялось разработке технических руководств по оцениванию инструментов КАТ (Green, Bock, Humphreys, Linn, & Reckase, 1984).

Адаптивное тестирование особенно подходит для использования в индивидуализированных программах обучения, упоминавшихся в главе 3. В этих случаях учащиеся проходят учебный предмет в удобном для себя темпе и могут поэтому выполнять

[1] Адрес дан в приложении Б. См. также Quan, Park, Sandahl, & Wolfe (1984) и Weiss, & Vale (1987).

значительно отличающиеся по трудности тестовые задания. Компьютеризованное тестирование позволяет прекращать проверку, как только ответы испытуемого дают достаточно информации для принятия решения об уровне овладения предметом. В настоящее время активно исследуются возможности применения компьютеризованного адаптивного тестирования в различных областях и соответственно разрабатываются технологии КАТ. В качестве одного из примеров можно привести разработанный совместно Службой тестирования в образовании и Советом колледжей компьютеризованный адаптивный тест для распределения поступивших в колледж студентов-первокурсников по группам для изучения английского языка и математики в соответствии с уровнем их подготовки по этим дисциплинам (Smittle, 1990; Ward, Kline, & Flaugher, 1986). Вследствие индивидуализированного подбора заданий этот тест почти не отнимает времени и позволяет сразу же получить оценку. Следовательно, его можно проводить в ходе регистрации поступивших и тут же распределять студентов по курсам или группам соответственно полученным результатам испытаний.

Еще одна важная область применения КАТ — крупномасштабные программы отбора и распределения персонала в промышленности, государственных учреждениях и армии. КАТ особенно хорошо подходит для этих целей, по меньшей мере, по трем причинам: 1) неуклонный рост потока кандидатов, которых необходимо испытать, и в связи с этим предотвращение тестирования очень больших групп, скапливающихся в одно время и в одном месте; 2) необходимость охватить широкий разброс уровня способностей и 3) лучшая защищенность теста, так как каждый кандидат получает разный набор заданий из большого банка заданий, хранящихся в памяти компьютера. Разработке КАТ версии Батареи профессиональной пригодности Вооруженных сил (*ASVAB*) предшествовало несколько лет поисковых исследований (McBride, & Martin, 1983; Moreno, Wetzel, McBride, & Weiss, 1984; Wiskoff, & Schratz, 1989). Постепенно разрабатываются КАТ версии всех важных групповых тестов, таких как Дифференциальные тесты способностей,[1] описанные в последнем разделе этой главы. Для многих практических приложений, равно как и для имеющих самостоятельное значение исследований природы и источников индивидуальных различий, КАТ дает бесспорные преимущества. Ясное и полезное изложение его перспектив для будущего тестирования можно найти в работе Embretson (1992).

Многоуровневые батареи

Общий обзор. В отличие от важнейших индивидуальных шкал и компьютеризованных адаптивных тестов в *традиционных* групповых тестах одни и те же задания предъявляются всем испытуемым, вне независимости от их индивидуальных ответов. По этой причине любой групповой тест должен включать задания относительно ограниченного диапазона трудности, пригодные для того конкретного возраста, класса или уровня способностей, для которых он предназначен. Чтобы обеспечить сравнимые меры интеллектуального развития в более широком диапазоне, была создана серия частично перекрывающихся многоуровневых батарей. Таким образом, любой конкретный человек обследуется только на подходящем для него уровне, а другие уровни могут использоваться для повторного тестирования того же человека в после-

---

[1] *DAT-Adaptive* соответствуют бланковой форме *DAT-Form V* (1981).

дующие годы или для получения сравнительных оценок разных возрастных групп. Частичное перекрытие последовательных батарей позволяет адекватно выявить нижнюю и верхнюю границы возможностей испытуемых, находящихся на краях своего возрастного диапазона или года обучения. Конечно, следует иметь в виду, что соответствие трудности задания и способности испытуемого, обеспечиваемое многоуровневыми батареями, в лучшем случае носит приблизительный характер. Более того, в отличие от индивидуализированных методик, реализующих принципы КАТ, это соответствие основывается на предварительной информации о тестируемых, такой как их возраст или класс, в котором они учатся, а не на их собственных ответах по тесту.

Многоуровневые батареи особенно полезны для использования в школах, где желательно достичь сопоставимости показателей на протяжении нескольких лет. По этой причине уровни батарей обычно описываются в терминах года обучения или класса школы. Большинство многоуровневых батарей обеспечивают достаточную степень непрерывности содержания или интеллектуальных функций, охватываемых батареей. Показатели повсюду выражаются в одной и той же шкале единиц. Для достижения непрерывности и сопоставимости показателей на всем протяжении диапазона измеряемой способности все больше и больше используются методы теории «задание — ответ» (*IRT*). В процессе стандартизации теста группам учащихся предъявляются частично перекрывающиеся уровни теста, с тем чтобы получить необходимые связующие данные. Нормативные выборки, обследуемые на разных уровнях, оказываются к тому же более эквивалентными, чем это имело бы место в случае независимо стандартизуемых тестах. Отдельные уровни охватывают от одного до трех классов школы. Суммарный же диапазон батареи в целом простирается от детей, посещающих детский сад, до студентов-первокурсников.

Большинство батарей дают общий стандартный показатель, соответствующий традиционному *IQ* в индивидуальных тестах. Некоторые батареи, наряду со стандартными показателями, предоставляют несколько типов норм, включая процентили, станайны или эквивалентные классы. В дополнение к суммарному общему показателю в большинстве батарей предусматриваются раздельные показатели по вербальным и количественным или лингвистическим и нелингвистическим заданиям. Такое разделение согласуется с данными о том, что выполнение конкретным человеком вербального и других типов субтестов может существенно расходиться, особенно на верхних уровнях.

Названия батарей также представляют определенный интерес. Для обозначения по существу одного и того же типа тестов используются такие термины, как «интеллект», «общие способности», «умственные способности», «умственная зрелость», «учебный потенциал» или «школьные способности». В словаре психометриста эти термины, фактически, являются синонимичными и взаимозаменяемыми. Примечательно, что в большинстве созданных в последнее время тестов или пересмотренных вариантов батарей термин «интеллект» заменен более специальными терминами. Такая замена объясняется тем, что термин «интеллект» приобрел слишком много побочных значений и его использование может привести к неправильному толкованию тестовых показателей. Многоуровневые батареи предназначены для выборочного измерения интеллектуальных умений и навыков, считающихся необходимыми для учебной деятельности. Главной целью таких батарей является оценка готовности индивидуума к обучению на каждой стадии образовательного процесса.

Типичные образцы батарей. Сущность и сферу действия современных многоуровневых батарей способностей можно проиллюстрировать на примере трех батарей, краткая характеристика которых дана в табл. 10–1. Эти батареи были выбраны из-за наличия свежих пересмотренных версий, высокого качества методов конструирования входящих в них тестов, а также объема и репрезентативности их выборок стандартизации. Еще одно достоинство выбранных батарей заключается в том, что их стандартизация проводилась параллельно со стандартизацией одной либо двух многоуровневых батарей тестов учебных достижений для тех же классов (о батареях тестов учебных достижений речь пойдет в главе 17). Благодаря проведению тестовых батарей обоих типов на одних и тех же выборках стандартизации появляется возможность установить соответствия между двумя множествами показателей. В результате эти два инструмента можно использовать совместно, что позволяет полнее исследовать развитие учащегося в процессе обучения и условия, влияющие на его развитие.

Надежность и валидность этих батарей широко исследовалась с помощью соответствующих методов. Коэффициенты надежности Кьюдера–Ричардсона как для общих показателей, так и для показателей по двум либо трем отдельным содержательным областям батарей, вычисленные по каждому уровню, в большинстве своем близки к 0,90. Ретестовые корреляции также высоки, что указывает на удовлетворительную устойчивость показателей. Корреляции со школьными отметками и с показателями тестов достижений свидетельствуют о хорошей прогностической валидности. Интеркорреляции частных показателей и результаты факторного анализа указывают на наличие выраженного общего фактора в каждой из полных батарей.

Типичное содержание тестов на различных уровнях. Доказано, что применение групповых тестов можно начинать с детей, посещающих детский сад и с первокласс-

Таблица 10–1

Типичные образцы многоуровневых батарей

Батарея	Охват классов	Число уровней	Нормирована совместно с
Тест школьных способностей Отиса–Леннона (*OLSAT*, 7-я ред.)	Д/с — 12	7	Серией Стэнфордских тестов достижений (9-я ред.)
Тест когнитивных способностей (*CogAT*, Form 5)	Д/с — 3 3–12	2 8	Тестами основных навыков штата Айова (д/с — 9-й кл.) Тестами достижений и умений (9 — 12) Тестами развития в обучении штата Айова (9–12-й кл.)
Тест когнитивных навыков (2-я ред., *TCS/2*)	2 — 12*	6	Калифорнийскими тестами достижений (5-я ред.) Комплексными тестами основных навыков (4-я ред.)

* Есть, кроме того, отдельная батарея — Элементарный тест когнитивных навыков (*Primary Test of Cognitive Skills* [*PTCS*]) — с иным набором тестов, предназначенных для уровня детского сада и 1-го класса.

ников. В дошкольном возрасте приходится использовать индивидуальные тесты для того, чтобы установить и поддерживать непосредственный контакт с ребенком, а также в силу необходимости предъявлять задания в устной и действенной форме, наиболее подходящей для маленьких детей. Однако уже детям 5–6 лет можно предъявлять отпечатанные тесты, при этом группы должны быть небольшими, до 10–15 человек. Но и при таком тестировании проводящий обследование должен по-прежнему уделять значительное внимание каждому ребенку, иначе он не сможет быть уверенным, что дети следуют инструкции; ему приходится следить, чтобы дети правильно переворачивали страницы тестовой тетради и соблюдали другие правила тестирования. При необходимости тестирующий вместе с одним-двумя помощниками может проводить обследование и с несколько бо́льшими группами.

Групповые тесты для элементарного уровня охватывают детский сад[1] и первые три класса начальной школы. В таких тестах каждый ребенок получает тетрадь с напечатанными картинками и схемами, составляющими задания теста; инструктирование ведется устно и обычно сопровождается показом. Часто включаются предварительные упражнения, в которых испытуемые пробуют выполнить один или два образца заданий, а тестирующий или его помощник проверяют ответы, чтобы быть уверенными, что инструкция понята правильно. Ребенок отмечает свои ответы в тестовой тетради цветным или простым карандашом. Большинство тестов требуют лишь умения правильно отметить картинку из данного набора изображений. Некоторые тесты требуют простой моторной координации, позволяющей, например, соединить линией две точки. Разумеется, тесты для элементарного уровня не требуют от обследуемых умения читать или писать.

Большинство многоуровневых батарей способностей включают тесты, пригодные для *элементарного уровня (primary level).* Типы тестовых заданий, используемых на этом уровне, приведены на рис. 10–3. Образцы этих заданий взяты из Теста школьных способностей Отиса—Леннона (*OLSAT*) и относятся к уровню *A*, пригодного для детей, посещающих детский сад. Результатом признания быстрого интеллектуального роста, происходящего в эти ранние годы, стало то, что в последней, седьмой редакции *OLSAT* предусмотрены четыре отдельных уровня (*A*, *B*, *C*, *D*) для воспитанников детского сада и учеников 1, 2 и 3-х классов соответственно. Эта редакция *OLSAT* обеспечивает бо́льшую дифференциацию по сравнению с более ранними редакциями этой батареи, да и по сравнению с другими многоуровневыми батареями тоже. На уровне *A* все инструкции даются тестирующим в устной форме. Ребенок реагирует на задания, закрашивая карандашом маленький кружок под выбранным в качестве ответа изображением, как показано на рис. 10–3, иллюстрирующем четыре из десяти типов заданий уровня А.

Для выполнения всего теста требуется около 75 мин. Он проводится в два этапа, на каждом из которых предусмотрен 5-минутный перерыв после первых 15–20 мин работы. Есть еще и Тренировочный тест (*Practice Test*) с похожими типами заданий и инструкциями, который может быть предложен в один из дней перед основным тестированием. Образцы заданий, показанные на рис. 10–3, являются относительно простыми и используются для того, чтобы познакомить детей с заданиями, которые им встретятся в самом тесте. Пояснения на рис. 10–3 представляют собой крайне сжатый

---

[1] В США детские сады (*kindergarten*) предназначены для воспитания и обучения детей в возрасте от 4 до 6 лет. — *Примеч. науч. ред.*

Классификация картинок: Отметьте картинку, не подходящую к остальным.

Фигурные аналогии: Поставьте метку под фигурой, которая должна находиться в пустом квадрате.

Следование указаниям: Отметьте число, находящееся прямо под «сердечком».

♥	★	D	●	5
E	●	★	H	♥
A	★	3	♥	2
♥	6	●	★	B

2 3 5 6

Последовательности картинок: Поставьте метку под картинкой, которая должна находиться в пустом квадрате.

Рис. 10-3. Образцы заданий, используемых в Тесте школьных способностей Отиса—Леннона (*OLSAT*, 7-я ред., уровень А)

вариант подробных устных инструкций и ясного описания содержания заданий, которыми сопровождается каждое из них. Реальные тесты имеют, к тому же, несколько иной формат, облегчающий понимание и помогающий маленьким детям удерживать внимание на заданиях. Например, листы и ряды изображений распознаются не только по номерам, но и по маленьким рисункам знакомых предметов, таких как чашка, ботинок или ножницы; кроме того, каждому ребенку дают маркер, чтобы он мог проследить ряд изображений, с которым должен работать.

Тесты для *уровня начальной школы (elementary school level)*[1], рассчитанные на учащихся 3–4-го класса и старше, весьма сходны как по своему содержанию, так и по построению. Поскольку учащиеся этой категории грамотны, преобладают тесты с вербальным содержанием, большинство тестов включают также арифметические задачи или иные числовые тесты. Кроме того, некоторые батареи имеют в своем составе тесты, не предполагающие умения читать, предназначенные для оценки тех же способностей к абстрактным рассуждениям у детей, не знающих английского языка, имеющих трудности с чтением или с усвоением других учебных навыков.

Типы заданий, соответствующих уровню начальной школы, проиллюстрированы на рис. 10–4. Эти задания являются типичными для промежуточных уровней Теста когнитивных способностей (*CogAT*). Как указано в табл. 10–1, *CogAT* включает два уровня, охватывающие период от детского сада до 3-го класса, и восемь уровней, приходящихся на период от 3 до 12-го класса. Тесты каждого уровня отпечатаны в отдельной тетради. Испытуемые, проходящие разные уровни теста, начинают и заканчивают работу заданиями, входящими в разные наборы. Тест построен таким образом, что большинство обследуемых выполняют задания среднего для них уровня трудности, что позволяет различить их наиболее эффективным образом.

Восемь уровней (от *A* до *H*) содержат одни и те же субтесты, сгруппированные в три батареи следующим образом.

Вербальная батарея — Классификация слов, Завершение предложений, Словесные аналогии.

Количественная батарея — Количественные отношения, Числовые ряды, Составление равенств.

Невербальная батарея — Классификация фигур, Фигурные аналогии, Анализ фигур. В этих субтестах не используются ни слова, ни числа, а только геометрические элементы и предметные изображения; их задания относительно слабо связаны со школьной программой.

Каждый субтест предваряется практическими упражнениями с подробными объяснениями. Кроме того, имеется Тренировочный тест, который может быть дан перед проведением основного теста. На рис. 10–4 показаны типичные задания шести из девяти субтестов такого теста, правда, с сокращенными и немного измененными инструкциями. По уровню трудности эти задания примерно соответствуют тем, которые предназначены для учащихся 4–6-х классов. В руководстве к *CogAT* рекомендуется предъявлять детям эти три батареи в три приема. Для большинства детей Невербальная батарея в отличии от Вербальной и Количественной батарей не является предсказателем достижений в учебе. Однако сравнительный анализ выполнения заданий по всем трем батареям может дать полезную информацию относительно специальных способностей или, напротив, «неспособностей» конкретного ребенка.

---

[1] В США начальная или, по-другому, элементарная школа охватывает первые 6–8 классов. – *Примеч. науч. ред.*

Классификация слов: Подумайте, чем похожи напечатанные жирным шрифтом слова, и найдите в нижнем ряду слово, которое к ним подходит.

добрый **дружелюбный** **помогающий**

A способный **B** активный **C** щедрый **D** симпатичный **E** сильный

Словесные аналогии: Подумайте, как связаны первые два слова из верхнего ряда, и укажите, какое слово из нижнего ряда точно так же связано с третьим.

корабль ⟶ **гавань** : **грузовик** ⟶

A шофер **B** шоссе **C** гараж **D** бензин **E** груз

Числовые ряды: Выведите правило, по которому построен расположенный ниже числовой ряд, и выберите из указанных чисел то, которое должно стоять в нем следующим.

3 2 1 3 2 1 ⟶

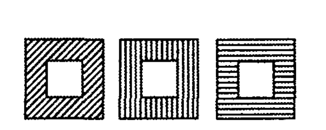
A 0 **B** 1 **C** 2 **D** 3 **E** 4

Составление равенств: Расположенные вверху числа и математические знаки можно объединить таким образом, что получится один указанных ниже ответов. Отметьте этот ответ.

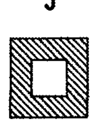
2 4 8 — —

J 0 **K** 2 **L** 4 **M** 6 **N** 10

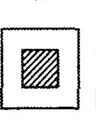
Классификация фигур: Первые три фигуры чем-то похожи. Найдите фигуру в правой части рисунка, которая имеет сходство с первыми тремя.



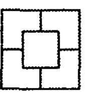
J



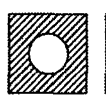
K



L



M



N



Фигурные аналогии: Догадайтесь, как связаны друг с другом первые две фигуры, и найдите справа фигуру, которая образует с третьей аналогичную пару.

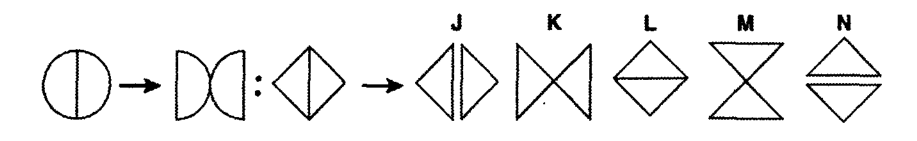


Рис. 10–4. Образцы некоторых типов заданий из Теста когнитивных способностей. Ответы отмечаются на отдельном бланке. Правильные ответы: *C, C, D, K, J, K*

(Из CogAT, Form 5, Practice Test for Levels A–H. Copyright © 1993 by The Riverside Publishing Company. Воспроизводится с разрешения)

Верхние уровни многоуровневых батарей, предназначенные для *учащихся средней школы (high school students)*,[1] в основе своей не отличаются от уровней, рассчитанных на учеников начальной школы, за исключением степени трудности. Эти уровни также пригодны для тестирования обычных, не отобранных специально групп взрослых, с самыми разными целями. Содержание тестов на этом уровне можно проиллюстрировать на примере заданий высшего уровня Теста когнитивных навыков (*TCS/2*). Каждый уровень этой батареи включает четыре теста:

Последовательности — уяснение и применение правила или принципа в отношении конфигурации или последовательности фигур, букв или чисел.

Аналогии — установление отношения внутри пары изображений и составление второй пары, демонстрирующей то же отношение; используются изображения сцен, людей, животных, предметов или графических символов.

Вербальное рассуждение — тестируется с помощью разнообразных типов заданий, среди которых установление существенных признаков предметов или понятий, классификация предметов по общим признакам, выявление отношений между двумя наборами слов или формулирование выводов из коротких отрывков текста.

Память — испытуемым предъявляют для заучивания набор искусственных слов (бессмысленных слогов) и через 25 мин (в это время проводятся другие тесты) проверяют их запоминание.

Здесь также есть Тренировочный тест, который дается за день или два до проведения основного теста. Примеры трех из четырех типов заданий приведены на рис. 10–5. В этой батарее одни и те же *типы заданий* из тестов «Последовательности», «Аналогии» и «Вербальное рассуждение» используются начиная с 4-го класса и далее, вплоть до 12-го класса, а одинаковые образцы заданий включены во все эти уровни. Два верхних уровня, соответствующие классам средней школы, выделены на основе установленной эмпирическим путем большей трудности их заданий.

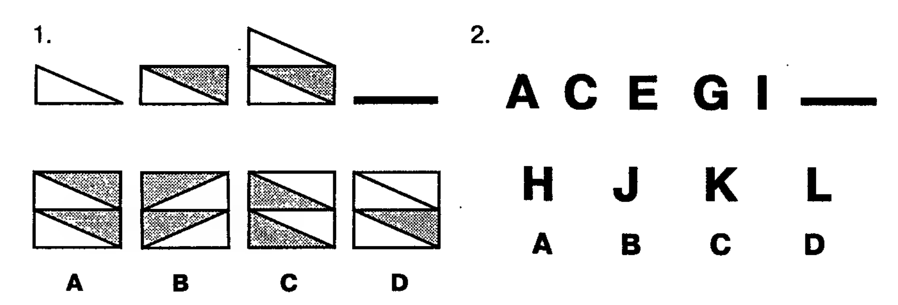
Отличительной особенностью батареи *TCS/2* является включение в нее теста памяти. Для вербального, невербального и мнемического тестов предусмотрено вычисление отдельных показателей. Эти области способностей были идентифицированы благодаря предварительному факторному анализу, результатами которого руководствовались при разработке и отборе заданий теста. Шкалирование выполнялось параллельно на всех уровнях в процессе стандартизации, с использованием методов теории «задание — ответ» (*IRT*, см. главу 7). С этой целью выборкам учащихся предъявлялись связующие тесты, содержащие задания из двух смежных уровней (*TCS/2, Technical Report*, 1993, p. 113–114). При создании батареи *TCS/2* были необычайно успешно применены методы *IRT* как для разработки тестовых заданий, так и для построения системы показателей. Вследствие этого ее показатели отражают не просто количество выполненных заданий, но и уровень трудности каждого из них.

Признание множественности способностей. Как уже отмечалось в первых разделах этой главы, существует явно выраженная тенденция к преодолению начального разрыва между тестированием единой, общей способности (*ability*) и измерением отдельных, относительно независимых способностей (*aptitude*). Преодоление этого

[1] То есть учащихся 9–12-х классов. — *Примеч. науч. ред.*

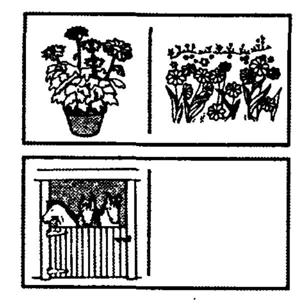
Последовательности

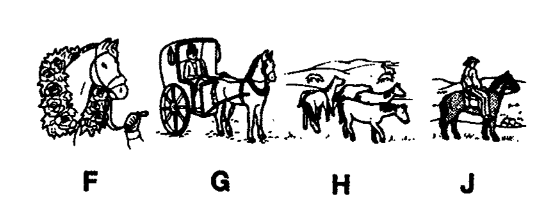
Разгадайте принцип организации каждой последовательности и выберите в нижнем ряду элемент, которым можно заполнить пробел.



Аналогии

Догадайтесь, как связаны друг с другом две картинки в верхних квадратах, и найдите справа картинку, которая образует с третьей аналогичную пару.





Вербальное рассуждение

1. Посмотрите на подчеркнутое слово: алфавит. Каким из расположенных ниже слов названо то, что должно всегда быть частью алфавита?

алфавит

A слова

B буквы

C цифры

D предложения

2. Учитывая информацию, содержащуюся в двух верхних предложения, решите, какое из приведенных ниже предложений должно быть истинным?

Большой Бен — часы в Англии.
Джуди осмотрела Большого Бена.

A Люди часто осматривают Большого Бена
B Многие часы в Англии — большие.
C Часы «Большой Бен» названы в честь какого-то человека.
D Джуди была в Англии.

Рис. 10–5. Образцы некоторых типов заданий, используемых в Тесте когнитивных навыков. Ответы отмечаются на отдельном бланке

(Из TCS/2 Practice Test, Levels 2–6. Copyright © 1992 by CTB/McGraw-Hill School Publishing Company. Воспроизводится с разрешения)

разрыва пошло с двух сторон, представленных сторонниками дискуссионных и поначалу казавшихся непримиримыми подходов к тестированию способностей. Отмеченная тенденция имеет параллель с тем, что происходило с индивидуальными тестами (см. главу 8). В данном случае первые многоуровневые батареи разрабатывались как групповые версии индивидуальных тестов интеллекта, хотя и имели обычно более узко определенную цель, а именно оценить академическую способность или готовность к переходу на следующий уровень образования. Со временем стало ясно, что единственный общий показатель иногда выгодно дополнять некоторыми более узкими показателями, что, собственно говоря, и было сделано во всех трех батареях, обсуждаемых в этом разделе.

В тесте Отиса—Леннона (*OLSAT*, редакция 1996 г.) было обращено внимание пользователей на то, что его суммарный показатель ограничен группой «вербально-учебных» (*verbal-educational*) способностей, и что в этой батарее не преследуется цель оценить «практически-технический» (*practical-mechanical*) компонент общего интеллекта. Более того, предусмотрено определение более узких дополнительных показателей в рамках вербального и невербального показателей батареи. Однако это разграничение обращено, в основном, к тем тестовым заданиям, которые требуют действия и не требуют употребления языка при ответах на тест. Таким образом, введенная дифференциация ориентирована, главным образом, на тестирование учащихся с ограниченным знанием английского языка. Впрочем, батарея позволяет еще получить показатели в шкале станайнов (в рамках возрастного уровня или школьного класса) для пяти кластеров тестов, выделенных внутри широких вербальных и невербальных категорий. Эти кластеры включают вербальное понимание, вербальное рассуждение, наглядно-образное рассуждение, символическое рассуждение и количественное рассуждение. В руководстве к батарее отмечается, что сравнение индивидуальных относительных результатов по этим кластерам может помочь в выявлении сильных и слабых сторон учащихся (OLSAT, 7th ed., Technical Manual, 1997).

В тесте когнитивных способностей (*CogAT, Form 5*, 1993) предусмотрены нормы не только для его общего суммарного показателя, но для показателей вербального, количественного и «невербального» (т. е. пространственного) рассуждения. Кроме того, бланки индивидуальных заключений по тесту содержат гистограммы для показателей в этих трех областях, а также для общего показателя. В руководстве к тесту особо привлекается внимание к полезности построения профилей с помощью таких гистограмм для предсказания учебных достижений. В инструкциях по интерпретации показателей неоднократно указывается на практическую важность рассмотрения профиля показателей индивидуума (см., например, Riverside, 2000, 1994, p. 44). Тем самым сделан еще один шаг вперед по пути признания множественности способностей.

В тесте когнитивных навыков (TCS/2, 1992) признание ценности данных о множественности способностей при оценке учебной деятельности выражено даже в более явной форме. Нормы по нему имеются не только для совокупных показателей по батарее, но также для каждого из четырех субтестов и для «невербального» блока в целом (субтестов «Последовательности» и «Аналогии»). Помимо этого, сама эта батарея разрабатывалась для оценки трех широких когнитивных свойств, установленных в результате факторно-аналитических исследований, а именно: Вербального рассуждения, Невербального рассуждения и Памяти. Таким образом, налицо определенное признание потребности в тестировании многих разнородных способностей, которое будет рассмотрено в следующем разделе.

Измерение множественных способностей

Традиционные тесты интеллекта, независимо от того, проводились ли они индивидуально или с группами, разрабатывались для получения одной-единственной, глобальной меры общего уровня когнитивного развития индивидуума, такой как *IQ*. Вскоре, однако, и практические, и теоретические результаты работы с ними привлекли внимание к ряду дифференцируемых способностей внутри того рыхлого конгломерата, который характеризовался с помощью первых тестов интеллекта. Это привело, с одной стороны, к конструированию отдельных тестов для измерения нескольких широко применимых способностей, а с другой — к уточнению определения и более полному пониманию того, что измеряли сами тесты интеллекта.

Росту интереса к измерению различных способностей способствовал ряд событий. Во-первых, происходило все большее осознание интраиндивидуальной вариации результатов выполнения тестов интеллекта. Грубые попытки сопоставить относительное положение индивидуума по разным субтестам или группам заданий многие годы предшествовали созданию батарей для оценки комплекса способностей или, короче, комплексных батарей способностей. Сами тесты интеллекта, однако, не предназначались для этой цели. Их субтесты или группы заданий часто были слишком ненадежны для того, чтобы можно было обоснованно проводить интраиндивидуальные сравнения. Кроме того, при конструировании тестов интеллекта задания или субтесты выбирались обычно таким образом, чтобы давать унитарную и внутренне согласованную меру. Поэтому при таком отборе все усилия прилагались к минимизации, а вовсе не к максимизации интраиндивидуальной вариации. Субтесты или задания, слишком слабо коррелирующие с остальной шкалой, как правило, исключались из теста. Хотя, вероятно, именно такие субтесты и задания как раз и следовало сохранить, если бы акцент ставился на дифференциации способностей. Вследствие такого способа конструирования большинства тестов интеллекта маловероятно, чтобы деятельность по выполнению этих тестов можно было значимо разделить более чем на две категории, таких как вербальная и невербальная или лингвистическая и количественная.

Дополнительным стимулом разработки комплексных батарей способностей послужило постепенное осознание того, что так называемые тесты общего интеллекта в действительности являются менее общими, чем первоначально предполагалось. Вскоре стало очевидным, что многие из этих тестов на самом деле служили средствами измерения вербального понимания. Определенные области, такие как область механических способностей, обычно в них не затрагивались, за исключением некоторых шкал действия и неязыковых шкал. По мере того как эти ограничения тестов интеллекта становились все очевиднее, психологи начали уточнять сам термин «интеллект». Одни из них предложили разграничивать «академический» и «практический» интеллект. Другие стали говорить об «абстрактном», «техническом» и «социальном» интеллекте. В дополнение к тестам интеллекта начали также конструировать тесты «специальных способностей» (*special aptitude*). Однако более тщательный анализ показал, что, вообще говоря, тесты интеллекта сами измеряют определенную комбинацию специальных способностей, таких как вербальные и числовые способности.

Мощный импульс развитию дифференциального тестирования способностей был придан также ростом активности психологов, работающих в сфере профконсультирования и планирования карьеры, а также занятых отбором и распределением персонала в промышленности и вооруженных силах. Самые ранние разработки специализиро-

ванных тестов для отбора конторских служащих, инженерно-технических работников и представителей ряда других профессиональных областей как раз и служат отражением таких интересов. Составление тестовых батарей для отбора абитуриентов, поступающих на медицинский, юридический, инженерно-технический, зубоврачебный и другие факультеты университетов, представляет собой аналогичную линию развития тестирования, сохранявшуюся в течение многих лет. Более того, ряд дифференциальных батарей способностей, таких как батареи, подготовленные военными и Управлением размещения и регулирования рабочей силы США (*U. S. Employment Service*), были прямым результатом деятельности специалистов, занимавшихся профотбором или распределением персонала.

Наконец, исследования структуры черт с помощью методов *факторного анализа*[1] обеспечили теоретическую основу для конструирования комплексных батарей способностей. Благодаря таким исследованиям появилась возможность систематически выявлять, классифицировать и определять разнообразные способности, связанные между собой весьма слабо, единственно за счет применения к ним общего термина «интеллект». Теперь тесты можно было отбирать таким образом, чтобы они являлись наилучшими из имеющихся средств измерения одного из факторов или одной из черт, выявляемых путем факторного анализа.

Дифференциальные тесты способностей. Одной из наиболее широко используемых комплексных батарей способностей являются Дифференциальные тесты способностей (*Differential Aptitude Tests* [*DAT*]). Впервые эта батарея тестов была издана в 1947 г. и в последующем периодически пересматривалась (5th ed., Form C, 1992). *DAT* предназначались главным образом для использования в профессиональном и образовательном консультировании учащихся 8–12-х классов. Пятая редакция *DAT* доступна пользователям в форме тестов двух уровней. Тесты 1-го уровня предназначены в основном для учащихся 7–9-х классов и взрослых с 7–9-летним образованием; тесты 2-го уровня предназначены для учащихся 10–12-х классов и взрослых с незаконченным средним (т. е. не менее чем 9-летним) образованием.

DAT включают в себя следующие восемь тестов: Вербальное рассуждение, Числовое рассуждение, Абстрактное рассуждение, Перцептивная скорость и точность, Пространственные отношения, Механическое рассуждение, Орфография и Словоупотребление. Образцы заданий из четырех тестов DAT показаны на рис. 10–6. Для предварительного ознакомления тестируемых с этой батарей имеется Тренировочный тест, охватывающий все восемь проверяемых областей. Существует также специальная форма *DAT* — Дифференциальные тесты способностей для оценки персонала и карьеры (*Differential Aptitude Tests for Personnel and Career Assessment*), — в которой каждый из восьми тестов сокращен и отпечатан в виде отдельной брошюры. Такая форма позволяет подбирать конкретные тесты для конкретных профессий и проводить их порознь.[2]

Подобно большинству основных современных тестов, *DAT* можно полностью проводить в компьютеризованной версии. В настоящее время опробуется более совре-

[1] Эта тема будет обсуждаться в главе 11.

[2] Инвентарь профессиональных интересов (*Career Interest Inventory*) был стандартизован вместе с 5-й редакцией *DAT*. Поэтому для решения задач профессионального и образовательного консультирования его можно проводить, подсчитывать и интерпретировать его показатели в сочетании с *DAT*.

Вербальное рассуждение

Выберите правильную пару слов и заполните пробелы в предложении. Первое слово пары вставляется в пробел в начале предложение, а второе — в пробел в конце предложения.

......нужен плавник, как птице......

A вода — перо **C** рыба — крыло

B акула — гнездо **D** дельфин — полет

E рыба — небо

Правильный ответ: **C**

Числовое рассуждение

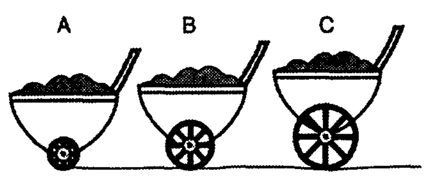
Какое число нужно подставить вместо R в этом примере на сложение?

7R +R	F	G	H	J	K
88	9	6	4	3	Ни одно из них

Правильный ответ: 9

Механическое рассуждение

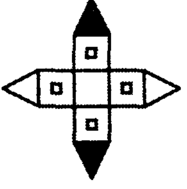
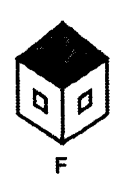
Какой груз будет легче везти по мягкому грунту?

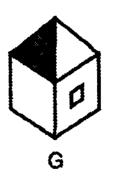


Правильный ответ: C

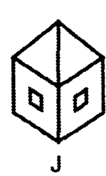
Пространственные отношения

Какая из следующих фигур могла бы дать изображенную слева развертку?







Правильный ответ: H

Рис. 10–6. Образцы заданий из Дифференциальных тестов способностей (5-я ред.)

менная разработка этой батареи в форме компьютеризованного адаптивного тестирования (*CAT*) — *DAT Adaptive*, доступная пользователям с 1987 г. Как при всяком адаптивном тестировании, каждый испытуемый получает здесь только те задания, которые соответствуют его уровню выполнения теста. В этой *CAT* версии использованы задания из более ранней версии *DAT* (*Form V*), которые были проанализированы на основе модели Раша — упрощенной, однопараметрической модели теории «задание — ответ» (см. главу 7).

За годы использования батареи *DAT* накоплена обширная коллекция данных о ее валидности, собранная как издателями, так и независимыми исследователями, применявшими *DAT*, в различных сферах профконсультирования и профотбора или включавшими эту батарею в исследовательские проекты. Большинство этих данных касаются прогностической валидности относительно достижений в учебных и профессиональных программах средней школы. Большинство коэффициентов валидности высоки, даже если вычислялись с интервалом в три года между тестированием и сбором данных о критериальной деятельности. В отношении дифференциального предсказания результаты оказались несколько менее ободряющими. Хотя, в общем, вербальные тесты имеют более высокие корреляции с курсами английского языка, а числовые тесты — с курсами математики, собранные данные свидетельствуют о существовании сильно выраженного общего фактора, лежащего в основе всякой успешной учебной деятельности. Тест «Вербальное рассуждение», например, имеет высокие корреляции с большинством учебных курсов. Главным образом по этой причине и был введен комплексный показатель *VR* + *NR* в качестве индекса академической способности.

Являясь суммой показателей по тестам «Вербальное рассуждение» (*VR*) и «Числовое рассуждение» (*NR*), этот индекс имеет корреляции в районе 0,70 — 0,80 с комплексным критерием учебных достижений. К индексу *VR* + *NR*, который является одним из показателей, регулярно включаемых в профиль *DAT* (см. рис. 4–6), имеются нормы. Существует также неполный вариант *DAT* — *Парциальная батарея (Partial Battery)*, содержащая только субтесты вербального (*VR*) и числового (*NR*) рассуждения, которую можно использовать в тех случаях, когда требуется лишь общий индекс академической способности.

С другой стороны, появляется все больше данных о том, что традиционные тесты «общего интеллекта» либо «академических способностей» — независимо от того, предназначены ли они для индивидуального или группового проведения, — дают существенные коэффициенты валидности относительно широкого множества образовательных и профессиональных критериев (L. S. Gottfredson, 1986a; Guion & Gibson, 1988; Pearlman et al., 1980; Schmidt, Hunter, Pearlman, & Shane, 1979). Такие тесты включают по существу тот же кластер когнитивных навыков и знаний, которые оцениваются показателем *VR* + *NR* из *DAT*. А это, как нетрудно заметить, свидетельствует о преодолении существовавшего ранее разрыва между тестами интеллекта и комплексными батареями с двух сторон. Тесты, подобные *DAT*, с самого начала придают повышенное значение использованию и интерпретации широких показателей, таких как *VR* + *NR*. В то же время в тестах первоначально общего характера все больше значения придается использованию и интерпретации показателей отдельных субтестов и анализу профиля. Что касается критических обзоров, посвященных *DAT*, см. работы Hattrup (1995) и N. Schmitt (1995).

Многоаспектная батарея способностей. В качестве примера инструмента, при создании которого достигнуто еще большее приближение к новой модели тестирования способностей, можно рассмотреть Многоаспектную батарею способностей (*Multidimensional Aptitude Battery* [*MAB*]). Впервые опубликованная в 1984 г., она затем была существенно обновлена в том, что касается процедур проведения, норм и руководства к тесту (Jackson, 1994b). *MAB* — это групповой тест, предназначенный для оценки тех же способностей, что и Пересмотренная шкала интеллекта взрослых Векслера (*WAIS-R*, см. главу 8). Он включает пять субтестов, составляющих Вербальную шкалу,[1] пять субтестов, организованных в Шкалу действия, и дает показатели — в единицах стандартного *IQ* — по Вербальной шкале (*V*), Шкале действия (*P*) и Полной шале (*Full Scale*). *MAB* пригодна для работы с подростками и взрослыми, однако эту батарею не рекомендуется применять для обследования лиц с задержкой психического развития или иными нарушениями умственной сферы, состояние которых могло бы помешать пониманию или соблюдению инструкций к тестам.

Десять субтестов *MAB*, имеющие за одним исключением те же названия, что и соответствующие субтесты *WAIS-R*, перечислены ниже:

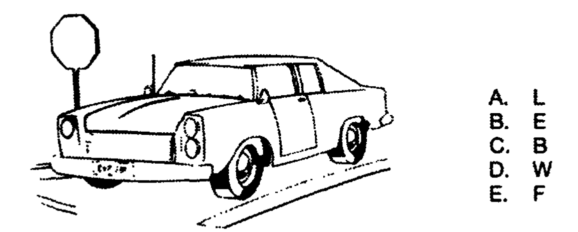
ВЕРБАЛЬНЫЕ СУБТЕСТЫ	СУБТЕСТЫ ДЕЙСТВИЯ
Осведомленность	Цифровые символы
Понимание	Недостающие детали
Арифметический	Пространственный
Сходства	Расположение картинок
Словарь	Складывание объекта

Субтест «Складывание кубиков» из *WAIS-R* был заменен в *MAB* субтестом «Пространственный» (*Spatial*). Создателям *MAB* пришлось проявить незаурядную изобретательность при разработке бланковых заданий, чтобы обеспечить измерение тех же функций, которые охватываются индивидуально проводимым тестом Векслера. Решить эту задачу было особенно трудно в отношении субтестов Шкалы действия. На рис. 10–7 приведены примеры простых, демонстрационных заданий из субтестов «Недостающие детали» и «Пространственного». Задача респондента во всех заданиях, входящих в каждый их этих субтестов, остается той же самой, как и в приведенных для иллюстрации заданиях. В субтесте «Недостающие детали» респондент должен решить, как называется недостающий элемент картинки, и затем выбрать первую букву этого названия среди предложенных вариантов. В субтесте «Пространственный» лишь одна из расположенных справа фигур могла бы быть получена простым поворотом на плоскости страницы фигуры, расположенной слева; все остальные варианты фигур, предлагаемых респонденту для выбора, требуют не только поворота, но и переворота исходной фигуры.

Пять Вербальных субтестов представлены в одном буклете, пять невербальных субтестов Действия — в другом. Каждый буклет начинается с задач для упражнения, иллюстрирующих типы заданий трех из пяти субтестов, а каждый субтест начинается с одного, двух или трех добавочных демонстрационных заданий. Общие и частные

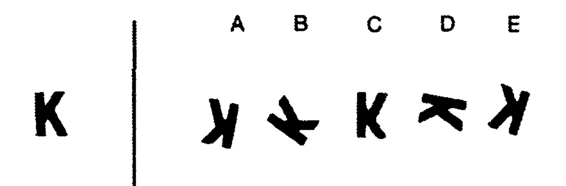
[1] Субтест «Повторение цифр» (*Digit Span*) из *WIAS-R* не имеет соответствия в *MAB*. Этот субтест было бы трудно воплотить в бланковой форме и к тому же он имеет самую низкую корреляцию с показателями Полной шкалы Векслера.

Недостающие детали. Выберите букву, с которой начинается слово, обозначающее пропущенную деталь картинки.



Правильный ответ — **Light** (Фара), поэтому в вариантах ответов следует зачеркнуть букву **А**.

Пространственный. Выберите одну фигуру справа от вертикальной линии, которая является той же самой, что и фигура слева от вертикальной линии. Искомую фигуру можно повернуть на плоскости, чтобы она выглядела как фигура слева; другие для этого пришлось бы еще и перевернуть.



Правильный ответ — **А**, поэтому в вариантах с ответом следует пометить букву **А**.

Рис. 10-7. Демонстрационные задания из двух тестов действия Многоаспектной батареи способностей (*MAB*)

(Copyright © 1983 by Douglas N. Jackson. Воспроизводится с разрешения)

инструкции для каждого субтеста приведены в руководстве, они могут даваться устно, в аудиозаписи или в виде текста на мониторе компьютера. Ответы фиксируются на отдельных бланках или, с помощью соответствующих устройств, на компьютере. Одна из последних версий *MAB* обеспечивает полностью компьютеризованное проведение и вычисление показателей батареи местным пользователем теста.

С помощью нормативной таблицы первичные показатели по каждому из 10 субтестов можно преобразовать в эквивалентные показатели единой равномерной шкалы (стандартные показатели с $M = 50$ и $SD = 10$). Суммы таких показателей по трем шкалам — *V*, *P* и *Full Scale* — рассматривают обычно как стандартные *IQ* ($M = 100$, $SD = 15$) в границах каждой из девяти возрастных групп, покрывающих возрастной диапазон от 16 до 74 лет. Кроме того, есть отдельные таблицы для нахождения в границах каждой из тех же девяти возрастных групп эквивалентных шкальных показателей, которые можно использовать при построении соответствующих возрасту профилей. Преимущество использования показателей из одной батареи по сравнению с показателями набора различных тестов заключается в том, что все тесты батареи были нормированы на одной и той же выборке стандартизации и, следовательно, допускают прямое сравнение результатов.

В целом, при разработке *MAB* были использованы психометрические методы, характеризующиеся высоким техническим качеством, и каждая стадия проекта поддерживалась интенсивными исследованиями, продолжавшимися более десяти лет.[1] Достойной упоминания особенностью *MAB* является ее эмпирическая состыковка с *WAIS-R*. Впервые было показано, что каждый субтест *MAB*, как и ее суммарные показатели по трем шкалам (*V*, *P* и *Full Scale*), имеют очень высокие корреляции с соответствующими показателями *WAIS-R*, полученные на неоднородной выборке 145 подростков и взрослых. Фактически, эти корреляции были столь же высоки, как и корреляции между показателями *WAIS* и *WAIS-R*, или даже выше, за двумя исключениями (субтесты «Цифровые символы» и «Пространственный»/«Складывание кубиков»). Следует отметить, что эти корреляции были получены несмотря на полную новизну заданий *MAB*, — в этих двух шкалах нет ни одного общего задания. На этом основании было выполнено линейное приравнивание показателей *MAB* и *WAIS-R*, проведенное на неоднородной выборке приравнивания, состоящей из 160 испытуемых в возрасте от 16 до 35 лет, которым предъявляли оба теста. Предварительное проведение приравнивания показателей этих двух тестовых батарей на выборках учащихся и пациентов психиатрических клиник показало, что такое градуирование распространимо на другие выборки тестируемых. Эти процедуры градуирования предлагают многообещающий способ выражения показателей впервые создаваемых тестов в единицах унифицированной шкалы, основанной на данных из большой, репрезентативной выборки стандартизации. Они представляют еще один шаг на пути к достижению такой важной цели, как построение национальных анкерных норм, обсуждавшихся в главе 3.

Если посмотреть с другой точки зрения, в этой батарее отчетливо выражена современная тенденция к иерархическим показателям. *MAB* дает полностью интерпретируемые показатели на уровне 10 субтестов, на более широком уровне Вербальной шкалы и Шкалы действия, и, наконец, обеспечивает получение общего суммарного показателя по полной батарее. Тем самым пользователь теста может проявить большую гибкость при выборе подходящего уровня показателей для своей специфической цели тестирования — условие, которое, как нам кажется, будет характеризовать тестирование способностей в XXI столетии.

---

[1] См. обзоры S. B. Reynolds (1989) и Silverstein (1989).

11 ПРИРОДА ИНТЕЛЛЕКТА

Все психологические тесты предназначены для измерения поведения. Поэтому подбор подходящих тестов и интерпретация результатов тестирования требуют знаний о человеческом поведении. Знание релевантных поведенческих исследований необходимо не только разработчику теста, но и его пользователю. В этой и следующей главах мы рассмотрим накопленные к настоящему времени знания о поведении, для оценки которого и предназначены тесты интеллектуальных умений и навыков. Нам предстоит разобраться в том, как психологические исследования способствуют пониманию 1) поведения, измеряемого тестами когнитивных способностей или «интеллекта», 2) источников индивидуальных различий в таком поведении и 3) предсказуемости такого поведения в последующем времени и в других условиях.

Прежде всего, следует отметить, что не имеющий строгого определения термин «интеллект» употребляется в огромном количестве значений, причем не только широкой публикой, но и представителями различных дисциплин, таких как биология, философия или педагогика (см. Sternberg, 1990), да и психологами, специализирующимися в разных областях или придерживающихся различных теоретических ориентаций (например, H. Gardner, 1983, 1993; Sternberg, 1985a, 1989 — см. также Brody, 1992; Lubinski & Benbow, 1995; Messick, 1992; H. Rowe, 1991). Самая первая демонстрация этого многообразия значений произошла в 1921 г, когда редактор «Журнала педагогической психологии» *(Journal of Educational Psychology)* предложил 17 ведущим исследователям сформулировать свои определения и понятия интеллекта («Intelligence...», 1921). Аналогичный опрос был проведен 65 лет спустя (Sternberg, & Detterman, 1986). Изучение этих публикаций, должно быть, представляет существенный теоретический интерес и могло бы обеспечить основу для глубокого обсуждения и, возможно, некоторого сближения конфликтующих позиций. В данном случае, однако, мы преследуем более ограниченную цель — выяснить, что нам следует знать о той специфической части человеческого интеллекта, которая оценивается посредством традиционных тестов интеллекта и обозначается символом *IQ*. *IQ* явно имеет более ограниченный смысл, чем тот, в котором термин «интеллект» употребляется при современном обсуждении этого конструкта (см. Anastasi, 1983c).

Значение IQ

В сознании широкой публики *IQ* не отождествляется с определенным типом показателя по конкретному тесту, а часто служит просто сокращенным обозначением интеллекта.[1] Такое употребление аббревиатуры *IQ* стало настолько преобладающим, что его нельзя больше игнорировать или осуждать как распространенное заблуждение. Несомненно, рассматривая количественное значение данного *IQ*, следует всегда точно указывать тест, при проведении которого этот показатель получен. Те или иные тесты интеллекта, дающие показатели в виде *IQ*, различаются и своим содержанием, и иными параметрами, влияющими на интерпретацию этого показателя. Некоторые из этих различий в тестах, объединяемых названием «тесты интеллекта», обсуждались в примерах, рассмотренных в предыдущих главах. Тем не менее не будет лишним еще раз рассмотреть превалирующие коннотации конструкта «интеллект» в том виде, как он символизируется *IQ*.

Во-первых, тестируемый интеллект следует рассматривать скорее как описательное, чем как объяснительное понятие. *IQ* — это форма выражения уровня способностей индивидуума в данный момент времени по отношению к имеющимся возрастным нормам. Ни один тест интеллекта не может указать на причины его результатов у конкретного человека. Отнесение неадекватного выполнения теста или обычной деятельности на счет «недостаточного интеллекта» есть тавтология, которая не только не продвигает нас в понимании умственного недостатка индивидуума, но фактически может замедлить исследование действительных причин такого недостатка в прошлом этого человека.

Тесты интеллекта, как и любые другие виды тестов, следует использовать не для навешивания ярлыков на людей, а для их лучшего понимания. Этот момент подчеркивался на протяжении многих лет во многих источниках — от работ психологов, специализирующихся в области индивидуальных различий, до официальных отчетов государственных комиссий (Hobbs, 1975a, 1975b; National Commission..., 1990). Широко разрекламированная книга (Herrnstein, & Murray, 1994) под названием «Гауссова кривая» (*The Bell Curve*) послужила еще большему укреплению разнообразных стереотипов и заблуждений, касающихся этнических и гендерных различий в выполнении тестов интеллекта, и только добавила путаницы и разногласий в отношении этой и без того сложной проблемы. Объективная и опирающаяся на факты трактовка относящихся к данной проблеме вопросов дана в отчете Специальной комиссии по интеллекту Американской психологической ассоциации (*American Psychological Association Task Force on Intelligence*, см. Neisser et al., 1996). Один из симпозиумов на съезде Американской психологической ассоциации в 1995 г. также был посвящен прояснению этих сложных вопросов (Steele, Chair, August, 1995). Чтобы повысить уровень функционирования конкретного человека до максимума, нужно исходить из того уровня, на котором он в данное время находится, а для этого необходимо оценить его сильные и слабые стороны и выработать соответствующий способ действий. Если тест на чтение показывает, что ребенок отстает в этом виде деятельности, мы ведь не останавливаемся на том, что навешиваем ему ярлык «плохо читает», и не даем ему невербальный

---

[1] Когда термин *IQ* (Коэффициент интеллекта) впервые вводился в обращение, он действительно имел отношение к типу показателя, а именно представлял собой отношение умственного возраста к хронологическому (см. главу 3).

тест, с тем чтобы спрятать этот недостаток за другим возможным достоинством. Вместо этого мы стараемся научить его нормально читать. Важной целью современного тестирования является к тому же содействие самопознанию и развитию личности. Данные тестирования все больше используются для того, чтобы помочь конкретным людям в планировании своего образования и профессиональной карьеры, а также в принятии оптимальных решений, непосредственно касающихся их жизни. Внимание, уделяемое эффективным способам сообщения тестовых результатов испытуемому, свидетельствует о растущем признании такого применения тестов.

Во-вторых, не следует забывать, что интеллект — это не единая, однородная способность, а композиция нескольких функций. Этим термином обычно обозначается сочетание способностей, необходимых для выживания и преуспевания в определенной культуре (Anastasi, 1986c). Следовательно, специфические способности, образующие эту композицию, а также их относительная значимость будут меняться в зависимости от времени и места. Для разных культур и в разные исторические периоды одной культуры понимание успешности в деятельности меняется. Изменение состава функций интеллекта можно видеть и на протяжении жизни одного человека от младенчества до взрослого состояния. Способность индивидуума будет с годами возрастать относительно тех функций, которым окружающая его культура или субкультура придают особое значение, и уменьшаться относительно тех функций, которым такого значения не придается.

Типичные тесты интеллекта, предназначенные для школьников или взрослых, измеряют в основном вербальные способности и, в меньшей степени, способности оперирования числами и другими абстрактными символами. Именно эти способности преобладают в школьном обучении. Большинство интеллектуальных тестов можно поэтому рассматривать как средство измерения способности к обучению или академического интеллекта. *IQ* является отражением предшествующих достижений в обучении и предиктором последующих. Поскольку функции, которыми овладевают в процессе получения образования, имеют первостепенное значение в современных культурах с передовыми технологиями, показатель по тесту академического интеллекта служит также эффективным предиктором успешной профессиональной и иной деятельности в таких культурах.

Вместе с тем множество других важных функций, таких как технические, двигательные, музыкальные и артистические способности, мотивационные, эмоциональные и диспозиционные (*attitudinal*) переменные, для измерения которых интеллектуальные тесты никогда не применялись, являются важными составляющими достижений во всех областях. В действительности же, некоторые психологи включают компоненты личности в свои определения интеллекта (например, H. Gardner, 1983). Аналогично этому, в исследованиях креативности выявляются когнитивные и личностные переменные, которые связаны с продуктивностью творческой деятельности. Все это, естественно, означает, что как индивидуальные решения, так и решения, принимаемые учреждениями, должны основываться на сопоставлении такого количества релевантных данных, которое только можно собрать. Принимать решения, основываясь исключительно на результатах тестов, особенно одного или двух, значит неправильно их использовать. Решения должны принимать люди, а тесты — всего лишь один из источников сведений, необходимых для принятия решений. Сами тесты не относятся к инструментам принятия решений.

Большинство наших знаний о том, что измеряют тесты интеллекта, мы получаем из практических исследований валидности тестов при предсказании образовательных и профессиональных достижений. На теоретическом уровне, в конце 1970-х гг. был отмечен сильный всплеск интереса к анализу конструкта «интеллект» в том виде как он измеряется тестами интеллекта (Humphreys, 1979; Resnick, 1976; Sternberg, & Detterman, 1979). Этот интерес оказался чрезвычайно устойчивым, захватывая разные области психологии и проникая через барьеры различных методологических подходов и теоретических ориентаций, что нашло свое отражение в продолжающейся серии публикаций (Detterman, 1985–1993; Sternberg, 1982–1989) и издании полной энциклопедии на эту тему (Encyclopedia of Human Intelligence, 1994).

Стремление понять, что же измеряют тесты интеллекта, связывалось не только с использованием стандартных статистических процедур, наподобие факторного анализа, но и с применением методов обработки информации к задачам, предъявляемым в интеллектуальных тестах (см. главу 5). Информационный подход сосредоточен на элементарных процессах, посредством которых тестируемый находит ответ на вопрос теста, а не на рассмотрении одной только правильности ответа. Этот тип анализа должен существенно помочь диагностическому использованию тестов и разработке обучающих программ, отвечающих специфическим индивидуальным потребностям.

Наследуемость и изменчивость[1]

Много недоразумений и споров возникло вследствие применения оценок наследуемости к показателям интеллектуальных тестов. В качестве примера можно привести известную статью А. Дженсена (Jensen, 1969), которая вызвала большой фурор и горячие споры, продолжающиеся и по сей день, то несколько утихая, то вспыхивая с новой силой. Хотя ее обсуждение шло по нескольким направлениям, а поднимавшиеся при этом вопросы были достаточно сложны, камнем преткновения для всех участников дискуссии оказалась интерпретация оценок наследуемости. Конкретно, коэффициент наследуемости показывает пропорциональный вклад генетических, или наследственных, факторов в общую изменчивость конкретного свойства или черты в данной популяции при существующих условиях. Например, утверждение, что наследуемость *IQ* по Стэнфорд–Бине среди учащихся американских городских средних школ составляет 0,70, означало бы, что 70 % дисперсии этого показателя может быть приписано наследственным различиям, а 30 % — влиянию среды.

Коэффициенты наследуемости вычислялись по разным формулам (см., например, Jensen, 1969; Loehlin, Lindzey, & Spuhler, 1975), но используемые для их расчета основные данные — это меры семейного сходства изучаемого признака. Наиболее распространенный метод состоит в использовании корреляций результатов интеллектуальных тестов у монозиготных и дизиготных близнецов. Также использовались корреля-

[1] Обсуждаемый в этом разделе вопрос касается лишь малой части обширной области исследований воздействия наследственности и среды на развитие поведения. Всестороннее рассмотрение генетических вопросов, включая критические оценки коэффициентов наследуемости, можно найти в работах Brauth, Hall, & Dooling (1991), Bronfenbrenner, & Ceci (1994), Horowitz (1994), Plomin, & McClearn (1993) и Plomin, & Reade (1991).

ции между монозиготными близнецами, воспитанными вместе, и между монозиготными близнецами, воспитанными порознь, в приемных семьях.

В интерпретации оценок наследуемости следует обратить внимание на ряд моментов. Во-первых, эмпирические данные относительно семейного сходства недостаточно точны, поскольку в них не учтен вклад средовых факторов. Например, имеются данные о том, что монозиготные близнецы живут в более сходной среде, чем дизиготные (Anastasi, 1958, p. 287–288; Koch, 1966), а среды растущих вместе сиблингов могут быть в психологическом плане совершенно различными (Daniels, & Plomin, 1985). Во-вторых, распределение пар близнецов по разным приемным семьям происходит отнюдь не случайным образом, как нужно было бы для проведения идеального эксперимента. Хорошо известно, что взятие ребенка на воспитание зависит и от особенностей малыша, и от характеристик приемной семьи. Следовательно, условия жизни близнецов в приемных семьях внутри каждой пары, по-видимому, будут иметь достаточно сходства, чтобы этим можно было объяснить хотя бы какую-то часть корреляции между их тестовыми показателями. Помимо того, есть некоторые основания утверждать, что данные о наследуемости, полученные близнецовым методом, нельзя обобщать на популяцию в целом, поскольку близнецы чаще подвергаются пренатальным травмам, приводящим к серьезным задержкам психического развития. Включение в выборку пар с сильной задержкой психического развития может заметно увеличить корреляцию результатов тестирования интеллекта близнецов (Nichols, & Broman, 1974).

Помимо сомнительности данных, используемых при вычислении коэффициентов наследуемости, последним присущи и другие серьезные ограничения (см. Anastasi, 1971; Hebb, 1970). Примечательно, что в первой части упомянутой статьи Дженсена (Jensen, 1969, p. 33–46) среди прочих назывались и они. Во-первых, понятие наследуемости применимо к популяциям, но не к отдельным индивидам. Например, при установлении этиологии психической задержки у конкретного ребенка коэффициент наследственности вряд ли окажет какую-либо помощь. Независимо от величины коэффициента наследуемости в данной популяции задержка психического развития у этого ребенка могла стать следствием дефектного гена (как при фенилкетонурии), пренатального повреждения головного мозга или крайней ограниченности опыта.

Во-вторых, коэффициенты наследуемости применимы только к той популяции, на которой в данное время они были получены, и любое изменение в наследственности или окружающих условиях может изменить этот коэффициент. Так, увеличение браков между кровными родственниками, например на изолированном острове, уменьшило бы дисперсию признаков, приписываемую наследственным факторам, и тем самым снизило бы коэффициент наследуемости; увеличение однородности среды, с другой стороны, уменьшило бы дисперсию признаков, относимую на счет средовых факторов, что привело бы к повышению коэффициента наследуемости. Кроме того, коэффициент наследуемости, рассчитанный на одной популяции, неприменим к анализу различий в выполнении теста двумя популяциями, такими как разные этнические группы.

В-третьих, наследуемость ничего не говорит о степени изменчивости признака. Даже если коэффициент наследуемости изучаемого признака в данной популяции равен 100 %, отсюда не следует, что влияние среды на формирование этого признака незначительно. Поясним этот момент следующим контрастным примером. Предположим, что в гипотетическом сообществе взрослых людей все питаются одинаково, т. е. каждый получает одну и ту же еду и в одинаковом количестве. В такой популяции

влияние особенностей питания на общую дисперсию здоровья и физического состояния будет нулевым, поскольку разницей в пище нельзя объяснить индивидуальные различия в здоровье и физическом развитии. Тем не менее если бы запасы продовольствия внезапно иссякли, все сообщество умерло бы от голода. Наоборот, улучшение качества пищи могло бы сказаться на общем улучшении здоровья членов этого сообщества.

Независимо от величины коэффициентов наследуемости, вычисляемых для *IQ* в разных популяциях, один эмпирический факт твердо установлен: *IQ* не является постоянной величиной и изменяется под воздействием окружающей среды. Некоторые основания для такого вывода рассматриваются в следующей главе, в связи с лонгитюдными исследованиями. В этих исследованиях был достигнут определенный прогресс в выявлении средовых условий, ускоряющих и замедляющих психическое развитие. Повышение и снижение *IQ* могут происходить как в результате случайных изменений в условиях жизни ребенка, так и под влиянием запланированного вмешательства со стороны его окружения. Важные изменения в составе семьи, резкое увеличение или снижение уровня семейного дохода, помещение в детский дом или обучение по программе подготовки в школу могут заметно увеличить или снизить *IQ*.

Интерес к систематическим программам развития интеллекта, возникший в конце 1970-х гг. в разных странах мира, сохраняется по настоящее время. Свидетельством тому является издание Международного информационного бюллетеня «Интеллект человека» *(Human Intelligence International Newsletter)* в период с 1980 по 1987 гг. Благодаря работе международной редакционной коллегии, этот информационный бюллетень раз в квартал освещал когнитивные исследования и приложение их результатов в сфере образования. Другой важный пример — десятилетняя программа в Венесуэле, утвержденная и систематически финансируемая правительством. Включающая в себя множество конкретных проектов по развитию «навыков мышления» (*thinking skills*), начиная от младенчества и кончая старостью, эта программа побудила ряд других стран к введению в действие аналогичных проектов (Collins and Mangieri, 1992; Greenwald, 1982, 1984; Herrnstein, Nickerson, Sánchez, & Swets, 1986; Nickerson, 1988; Spitz, 1986; Sternberg, 1986).

Исследования результатов спланированного вмешательства на уровне младенчества и дошкольного детства будут рассмотрены в главе 12. Стоит, однако, отметить увеличение объема данных, демонстрирующих эффективность такого вмешательства на более поздних стадиях жизни. Хотя и менее масштабные, чем ориентированные на дошкольников, программы для детей школьного возраста также дали обнадеживающие результаты (Bloom, 1976; Brown, & Campione, 1986; Campione, & Brown, 1987; Jacobs, & Vandeventer, 1971; Olton, & Crutchfield, 1969; Resnick, & Glaser, 1976). Некоторые исследователи работают с еще более взрослым контингентом — студентами колледжей и профессиональных школ; и они тоже сообщают о значительном улучшении как академических достижений, так и показателей тестов академических способностей у студентов, включенных в программы вмешательства (Bloom, & Broder, 1950; Whimbey, 1975, 1977, 1980).В исследованиях на лицах пожилого возраста также получены доказательства эффектов научения и переноса у участников программ обучающего вмешательства (Willis, Blieszner, & Baltes, 1981). Другие исследователи работали с обучаемыми умственно отсталыми детьми и подростками, и опять-таки добились существенных улучшений (Babad, & Budoff, 1974; Budoff, & Corman, 1974; Feuerstein,

1980; Feuerstein et al., 1987; Hamilton, & Budoff, 1974; Rand, Tannenbaum, & Feuerstein, 1979).[1]

Эти программы обеспечивают обучение широко применяемым когнитивным навыкам, стратегиям решения задач (*problem-solving*) и эффективным приемам учения. Особый интерес представляют программы, в которых сделан акцент на развитие текущего самоконтроля и самокритики как условий эффективной деятельности (Flavell, 1979; Owings, Petersen, Bransford, Morris, & Stein, 1980; Whimbey, 1975). Оценка человеком своего уровня деятельности и осознание того, что ему понятно и что непонятно, представляет собой первый важный шаг к улучшению своих результатов. Все еще слишком часто неуспевающий ученик не способен отличить подлинное понимание от неточного или поверхностного. Мы располагаем данными о том, что детям с трудностями в обучении (*learning disabilities*) особенно не достает самокритики и способности осуществлять текущий контроль своей познавательной деятельности (Kotsonis, & Patterson, 1980).

Другие примеры тех видов когнитивных умений и навыков, которым обучают в этих программах интеллектуального развития, приводились в главе 1. Там такое обучение широко применимым когнитивным навыкам противопоставлялось натаскиванию в выполнении узко ограниченных тестовых заданий. Как отмечалось в этой связи, решающий вопрос, требующий ответа при оценивании программ интеллектуального развития, касается степени переноса или распространяемости эффектов обучения за пределы того содержания и той обстановки, которые характеризовали ситуацию обучения. Связанный с ним вопрос относится к прочности достигнутого улучшения.

Еще один предмет для рассмотрения — время, требуемое уже немаленькому ребенку или взрослому, чтобы накопить объем знаний, составляющих неотъемлемую часть интеллекта и влияющих на готовность их обладателя к усвоению более сложного материала. Становится все больше доказательств в пользу того, что схемы решения задач и понятия, за исключением самых элементарных уровней, связаны со специфическими предметными областями. Так, навыки решения задач тесно связаны с хранящимся в памяти организованным содержанием, накопленным индивидуумом в конкретной области знаний (Bransford, Sherwood, Vye, & Rieser, 1986; Brown, & Campione, 1986; Glaser, 1984; Larkin, McDermott, Simon, & Simon, 1980a; Neimark, 1987; Resnick, & Neches, 1984; Richardson, Angle, Hasher, Logie, & Stoltus, 1996). Хотя взрослый, опытный человек, вооруженный эффективными методами учения, может создать этот необходимый запас знаний быстрее, чем если бы он был ребенком, вряд ли стоит рассчитывать на то, что это произойдет за время его участия в короткой обучающей программе. Чем старше человек, тем больший пробел в знаниях придется ему заполнять. Неспособность понять это может привести к разочарованию и ослабить веру в эффективность всех таких обучающих программ.

Мотивация и интеллект

Хотя классификация тестов на отдельные категории привычна и общепризнанна, следует помнить, что любое такое различение в значительно мере поверхностно. При интерпретации тестовых показателей личность и способности невозможно развести.

[1] Что касается критических оценок подхода Фейерстейна (Feuerstein), см. Anastasi (1980) и Blagg (1991).

На выполнении конкретным человеком теста способностей, так же как и на его учебе, работе или ином виде деятельности, сказываются его стремление к достижениям, настойчивость, система ценностей, умение освободиться от затруднений эмоционального порядка и другие характеристики, традиционно связываемые с понятием «личность».

Имеет место растущее признание роли мотивации учащихся в школьном обучении (Bloom, 1976, chap. 4; Budoff, 1987; Feuerstein et al., 1987; J. G. Nichols, 1979; Renninger, Hidi, & Krapp, 1992; R. E. Snow, 1989). Интересы и аттитюды индивидуума, представление о себе как ученике влияют на его открытость сообщаемой на уроке информации и желание хорошо ее усвоить, на его внимание к учителю и время, уделяемое им выполнению задания. И мы располагаем данными, что эти индивидуальные реакции существенно связаны с достижениями в учебе (Baron, 1982; Dreger, 1968; J. McV. Hunt, 1981).

На более базисном уровне отмечается растущее согласие по поводу того, что способности больше не могут исследоваться независимо от аффективных переменных (Anastasi, 1985 b, 1994; Izard, Kagan, & Zajonc, 1989; Kanfer, Ackerman, & Cudeck, 1989, Part IV; Moore, & Isen, 1990; Saklofske, & Zeidner, 1995; Salovey, & Sluyter, 1997; R. E. Snow, 1992; Spaulding, 1994; Sternberg, & Ruzgis, 1994).

Воздействие вре́менных эмоциональных состояний на текущую деятельность человека надежно установлено. Еще более важным является кумулятивное воздействие черт личности на направление и степень интеллектуального развития индивидуума. Подтверждающие данные получены в исследованиях разного рода, включая длительные лонгитюдные (Eichorn, Clausen, Haan, Honzik, & Mussen, 1981) и более современные проекты, с использованием методов моделирования структурными уравнениями для выявления причинных связей (Shavelson, & Bolus, 1982). Такие исследования снабжают нас данными о том, что предсказание последующего интеллектуального развития индивидуума можно существенно улучшить, объединяя информацию о мотивации и аттитюдах с показателями тестов способностей.

Один из путей влияния мотивации и других аффективных переменных на развитие способностей связан с суммой времени, отводимой человеком на определенную деятельность относительно других возможных занятий, конкурирующих с ней за внимание с его стороны. На основе 25-летнего изучения мотивации достижения Дж. Аткинсон и его коллеги (Atkinson, 1974; Atkinson, O'Malley, & Lens, 1976) составили подробную схему взаимосвязей способностей, мотивации и факторов окружающей среды. Ключевым в этой схеме является понятие «времени на задачу» (*time-on-task*), т. е. времени, уделяемому индивидуумом какому-то одному виду деятельности, например изучению или выполнению связанных с работой функциональных обязанностей. Мотивация влияет как на эффективность выполнения задачи, так и на затраченное на нее время относительно других занятий. Уровень выполнения зависит от соответствующих способностей индивидуума и от эффективности, с какой он использует эти способности для выполнения поставленной задачи. Конечное достижение, или результат, отражает совместное действие уровня выполнения задачи и затраченного на нее времени.

Другой важный компонент схемы Аткинсона имеет отношение к долговременному, кумулятивному воздействию выполнения задачи на собственное когнитивное и мотивационное развитие индивидуума. Эта ступень схемы отображает цепь обратной связи, направленную к собственным свойствам и чертам индивидуума, и осуществля-

емое через нее влияние должно проявляться как в будущих тестовых показателях, так и в результатах реальной деятельности. Прогностическая ценность схемы Аткинсона подтверждена результатами машинного моделирования и эмпирического анализа данных лонгитюдных исследований учащихся средней школы (Atkinson, 1974; Atkinson et al., 1976; Lens, Atkinson, & Yip, 1979).

Эффект чистого «времени на задачу» усиливается контролем внимания. Чему конкретно человек уделяет внимание, насколько он способен сосредоточиться и сколько времени может удерживать его на предмете, — все это влияет на когнитивный рост данного человека. Избирательность внимания ведет к избирательному научению, — и этот выбор будет различаться у разных людей, находящихся в одинаковой непосредственной обстановке. Более того, такое избирательное научение может влиять на относительное развитие различных способностей и через это на формирование разных структур черт индивидуума (Anastasi, 1970, 1983a, 1986b). По существу, отдельные аспекты контроля внимания усиливают эффект времени, уделяемого значимым занятиям, и тем самым увеличивают его воздействие на развитие способностей.

Отношения между личностью и интеллектом реципрокны. Не только качества личности влияют на интеллектуальное развитие, но и интеллектуальный уровень может влиять на развитие личности. Свидетельствующие о такой связи данные были получены Плантом и Миниумом (Plant, & Minium, 1967). Используя данные из пяти лонгитюдных исследований молодых людей, закончивших колледжи, авторы отобрали в каждой выборке по результатам интеллектуальных тестов 25 % студентов, лучше всех выполнивших тесты, и 25 %, выполнивших тесты хуже всех. Полученные контрастные группы затем сравнивались по результатам ряда личностных тестов, ранее предъявлявшихся одной или более выборкам и включавших измерение аттитюдов, ценностей, мотивации, межличностных и других некогнитивных черт. Анализ этих данных показал, что «более способные» группы по сравнению с «менее способными» значительно сильнее подвержены «психологически позитивным» изменениям личности.

Результат, которого конкретный человек добивается в развитии и использовании своих способностей, зависит от особенностей эмоциональной регуляции, характера межличностных отношений и сложившихся представлений о самом себе (так называемой Я-концепции). В Я-концепции особенно явно проявляется взаимное влияние способностей и черт личности. Успехи ребенка в школе, на игровой площадке и в других ситуациях помогают ему формировать представление о самом себе, а его Я-концепция на данном этапе влияет на последующее выполнение им своих ролей и т. д. по спирали. В этом смысле Я-концепция действует как своего рода личное самоосуществляющееся пророчество.

В последние годы возрос интерес к изучению роли аффективных факторов в развитии младенцев. В ряде исследований были установлены существенные корреляции между оценками (*ratings*) поведения младенцев по личностным переменным и последующим когнитивным развитием, оцениваемым с помощью таких инструментов, как *WISC-R* и шкала Стэнфорд—Бине (Birns, & Golden, 1972; R. B. McCall, 1976; Palisin, 1986; Yarrow, & Pedersen, 1976). В общем, младенцы, демонстрирующие положительные эмоции, активный интерес и быстроту реагирования в тестовой ситуации, все же быстрее научаются и быстрее продвигаются в своем когнитивном развитии в результате раннего приобретения более богатого опыта. Кроме того, они, по-видимому, более благосклонно относятся к последующим учебным занятиям, включающим взаимодействие со взрослым в ходе решения задач с заданной целью. Дополнительное пре-

имущество возникает из того влияния, которое такое поведение малышей оказывает на социальное поведение ухаживающих за ними взрослых, что, в свою очередь, увеличивает благоприятные для ребенка возможности научиться чему-то новому (Haviland, 1976; Wilson, & Matheny, 1983).

Если говорить более конкретно, исследования мотива овладения средой у младенцев выявили некоторые многообещающие связи с более поздними замерами интеллектуальной компетентности. Направленное на овладение средой поведение младенца включает наблюдение, исследование и манипулирование элементами ближайшего окружения. Этот мотив, по природе своей, должен быть главным «помощником» когнитивного развития, и действительно, в публикациях экспериментального характера приводится все больше доказательств в пользу такого утверждения (Hrncir, Speller, & West, 1985; White, 1978; Yarrow et al., 1984; Yarrow et al., 1983). Фактически, некоторые из этих результатов наводят на мысль о том, что ранние признаки мотивации овладения средой, возможно, являются лучшим предиктором последующей интеллектуальной компетентности ребенка, чем ранние замеры самой компетентности. Изучение младенцев ведет к сближению исследований аффективного и когнитивного развития. Возможно, это приведет в конечном счете к более интегрированному использованию аффективных и когнитивных данных в интерпретации результатов тестов на любом возрастном уровне.

Факторный анализ интеллекта

Психологические исследования, цель которых — идентификация психических черт, выросли из интереса ученых к природе и структуре человеческого интеллекта.[1] Такие исследования начинаются с вычисления интеркорреляций показателей, полученных на выборке испытуемых по широкому набору тестов способностей. Затем корреляционная матрица подвергается дальнейшему математическому анализу с целью выявления общих факторов или черт на множестве тестов. Имеющиеся для достижения этой цели разнообразные методы объединены под общим названием *факторного анализа.*

Факторная матрица. Основная цель факторного анализа (ФА) — упростить описание данных посредством сокращения числа необходимых переменных или, иначе говоря, сократить размерность пространства описания данных. Так, если установлено, что пяти факторов достаточно для объяснения всей общей дисперсии в батарее из 20 тестов, то в большинстве случаев исходные 20 показателей без существенной потери информации можно заменить пятью новыми показателями. На практике обычно из совокупности исходных тестов сохраняют те, которые дают лучшие меры каждого из факторов.

Факторный анализ, независимо от используемых методов, начинается с обработки таблицы интеркорреляций, полученных на множестве тестов, известной как корреляционная матрица, а заканчивается получением факторной матрицы, т. е. таблицы, показывающей вес или нагрузку каждого из факторов по каждому тесту. Табл. 11–1 представляет собой гипотетическую факторную матрицу, включающую всего два фак-

[1] История этого вопроса затрагивается в Anastasi (1984b).

Таблица 11–1

Гипотетическая факторная матрица

Тест	Фактор I	Фактор II
1. Словарный	0,74	0,54
2. Аналогии	0,64	0,39
3. Завершение предложений	0,68	0,43
4. Восстановление порядка слов в предложении	0,32	0,23
5. Понимание прочитанного	0,70	0,50
6 Сложение	0,22	-0,51
7. Умножение	0,40	-0,50
8. Арифметические задачи	0,52	-0,48
9. Составление уравнений	0,43	-0,37
10. Завершение числовых рядов	0,32	-0,25

тора. Факторы перечисляются в верхней строке таблицы от более значимого к менее значимому, а их веса в каждом из 10 тестов даны в соответствующих столбцах.

Разработано несколько различных методов разложения множества переменных на общие факторы. Еще в начале века Карл Пирсон (Pearson, 1901) показал способ решения задачи такого типа, а Чарльз Спирмен (C. Spearman, 1904, 1927) заложил основы современного факторного анализа. Т. Келли (T. L. Kelly, 1935) и Л. Тёрстоун (L. L. Thurstone, 1947) в Америке и С. Берт (C. Burt, 1941) в Англии много сделали для усовершенствования этого метода. Альтернативные методы, модификации и усовершенствования ФА разрабатывались многими авторами. Наличие быстродействующих вычислительных машин ведет к принятию более тонких и, соответственно, трудоемких методов ФА. Несмотря на разницу в исходных постулатах, большинство этих методов дает сходные результаты. Для детального знакомства с методами ФА читатель может обратиться к учебникам Comrey, & Lee (1992) или Loehlin (1992). Краткий и простой обзор основных понятий и методов ФА можно найти в книгах Kim, & Mueller (1978a, 1978 b) и P. Kline (1993).

Рассмотрение математических основ или вычислительных процедур ФА не входит в задачи этой книги. К счастью, для понимания результатов ФА не обязательно владеть его специальной методологией. Даже без знания того, как были вычислены факторные нагрузки, можно понять, каким образом следует использовать факторную матрицу для идентификации и интерпретации факторов. Тем не менее, чтобы с пользой читать публикации, посвященные факторно-аналитическим исследованиям, знакомство с некоторыми понятиями и терминами ФА не помешает.

Оси координат. Принято представлять факторы геометрически в виде осей координат, относительно которых каждый тест может быть изображен в виде точки. Рис. 11–1 поясняет эту процедуру. На этом графике каждый из 10 тестов, приведенных в табл. 11–1, отображен в виде точки относительно двух факторов, которые соответствуют осям I и II. Так, тест 1 представлен точкой с координатами 0,74 по оси I и 0,54 по оси II. Точки, представляющие остальные 9 тестов, построены аналогичным способом, с использованием значений весов из табл. 10–1. Все веса по фактору I положи-

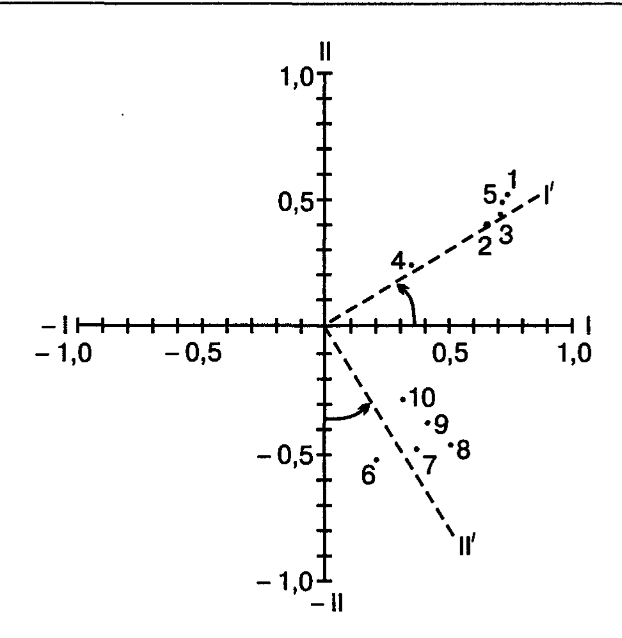


Рис. 11–1. Гипотетическое факторное отображение, показывающее веса двух групповых факторов по каждому из 10 тестов

тельны, веса по фактору II как положительны, так и отрицательны, что также отражено на рис. 11–1, где тесты с 1-го по 5-го образуют кластер в одной части координатной плоскости, а тесты с 6-го по 10-го — в другой.

В этой связи следует заметить, что положение осей координат не фиксировано данными. Исходная таблица корреляций определяет лишь положение тестов (т. е. точек на рис. 11–1) *относительно друг друга.* Те же точки можно нанести на плоскость с любым положением координатных осей. По этой причине при проведении факторного анализа обычно вращают оси до тех пор, пока не получают наиболее приемлемого и легко интерпретируемого отображения. Эта процедура вполне обоснованна и в чем-то похожа на измерение долготы, скажем, не от гринвичского меридиана, а от проходящего через Чикаго.

На рис. 11–1 полученные после вращения оси I′ и II′ показаны пунктирными линиями[1]. Это вращение выполнено в соответствии с предложенными Тёрстоуном критериями *положительного многообразия и простой структуры.* Первый предполагает вращение осей до положения, при котором исключаются все значимые отрицательные веса. Большинство психологов считают отрицательные факторные нагрузки логически несоответствующими тестам способностей, так как такая нагрузка означает, что чем выше оценка индивидуума по специфическому фактору, тем ниже будет его результат по соответствующему тесту. Критерий простой структуры, в сущности, означает, что каждый тест должен иметь нагрузки по как можно меньшему числу факто-

[1] Читатель, вероятно, заметил, что полученную в результате вращения ось II′ следовало бы обозначить как –II′, чтобы привести в соответствие с неповернутой осью –II. Однако какой из полюсов оси выбрать в качестве положительного или отрицательного, дело произвольное. В нашем примере полученная вращением ось II′ была «перевернута», с тем чтобы избавиться от отрицательных весов.

Таблица 11–2

Факторная матрица после вращения

Тест	Фактор I′	Фактор II′
1. Словарный	0,91	–0,06
2. Аналогии	0,75	0,02
3. Завершение предложений	0,80	0,00
4. Восстановление порядка слов в предложении	0,39	–0,02
5. Понимание прочитанного	0,86	–0,04
6. Сложение	–0,09	0,55
7. Умножение	0,07	0,64
8. Арифметические задачи	0,18	0,68
9. Составление уравнений	0,16	0,54
10. Завершение числовых рядов	0,13	0,38

(По данным, представленным на рис. 11–1)

ров.[1] Выполнение обоих критериев дает факторы, которые можно наиболее легко и однозначно интерпретировать. Если тест имеет высокую нагрузку по одному фактору и не имеет значимых нагрузок по другим факторам, мы можем кое-что узнать о природе этого фактора, изучив содержание данного теста. Напротив, если тест имеет средние или низкие нагрузки по шести факторам, то он мало что скажет нам о природе любого из них.

На рис. 11–1 хорошо видно, что после вращения осей координат все вербальные тесты (1–5) располагаются вдоль или очень близко к оси I′, а числовые тесты (6–10) тесно группируются вокруг оси II′. Новые факторные нагрузки, измеренные относительно повернутых осей, приведены в табл. 11–2. Читатель может легко проверить значения этих факторных нагрузок, изготовив из бумаги «масштабную линейку» со шкалой единиц, соответствующей масштабу координатных осей. С помощью этой линейки можно измерить длину отрезков, соответствующих проекциям точек (тестов) на повернутые оси координат. Факторные нагрузки в табл. 11–2 не имеют отрицательных значений, за исключением пренебрежимо малых величин, явно относимых к ошибкам выборки. Все вербальные тесты имеют высокие нагрузки по фактору I′ и практически нулевые — по фактору II′. Числовые тесты, напротив, имеют высокие нагрузки по фактору II′ и пренебрежимо низкие — по фактору I′. Таким образом, вращение координатных осей существенно упростило идентификацию и называние обоих факторов, а также описание факторного состава каждого теста. На практике число факторов часто оказывается больше двух, что, разумеется, усложняет их геометрическое представление и статистический анализ, но не изменяет существа рассмотренной процедуры.

Некоторые исследователи руководствуются теоретической моделью как принципом вращения осей. Кроме того, принимается в расчет неизменность, или подтверж-

---

[1] Этот критерий требует, чтобы по некоторым факторам тесты имели нагрузки, значимо не отличающиеся от нуля. Такое требование можно теперь проверить эмпирически, используя доступные статистические процедуры для нахождения стандартной ошибки факторных нагрузок (Cudeck, & O'Dell, 1994).

дение одних и тех же факторов в независимо выполненных, но сравнимых исследованиях. В настоящее время факторный анализ все чаще используется в роли подтверждающего, чем исследовательского метода. Нередко его сочетают с моделированием структурными уравнениями (см. главу 5) для оценивания теоретически сформулированной модели вклада различных переменных в выполнение задачи (см., например, Loehlin, 1992).

Интерпретация факторов. Получив после процедуры вращения факторное решение (или, проще говоря, факторную матрицу), мы можем переходить к интерпретации и наименованию факторов. Этот этап работы скорее требует психологической интуиции, нежели статистической подготовки. Чтобы понять природу конкретного фактора, нам ничего не остается, как изучить тесты, имеющие высокие нагрузки по этому фактору, и попытаться обнаружить общие для них психологические процессы. Чем больше оказывается тестов с высокими нагрузками по данному фактору, тем легче раскрыть его природу. Из табл. 11–2, к примеру, сразу видно, что фактор I′ вербальный, а фактор II′ числовой.

Приведенные в табл. 11–2 факторные нагрузки отображают к тому же корреляцию каждого теста с фактором.[1] Напомним, что эта корреляция есть не что иное, как факторная валидность теста (глава 5). По табл. 11–2 можно, к примеру, определить, что факторная валидность словарного теста как средства измерения вербального фактора равна 0,91. Факторная валидность теста на сложение относительно числового фактора равна 0,55. Очевидно, что первые 5 тестов имеют пренебрежимо малую валидность как средства измерения числового фактора, а последние 5 — практически нулевую валидность в качестве мер вербального фактора.

Факторная композиция теста. Одна из основных теорем ФА гласит: полная дисперсия теста равна сумме дисперсий, обусловленных действием общих (разделяемых с другими тестами) и специфических (встречающихся только в одном таком тесте) факторов, плюс дисперсия ошибок.

Мы уже сталкивались с дисперсией ошибок при анализе показателей тестов (глава 4). Если, к примеру, коэффициент надежности теста равен 0,83, то это значит, что 17 % дисперсии показателей по этому тесту составляет дисперсия ошибок (1,00 – 0,83 = 0,17). При помощи факторного анализа можно провести более тонкий анализ источников дисперсии, влияющих на выполнение того или иного теста.

Рассмотрим два гипотетических теста, информация о которых представлена в табл. 11–3. В ней для каждого теста указаны его факторные нагрузки по Вербальному *(V)*, Числовому (*N*) и Логическому *(R)* факторам, а также коэффициенты надежности этих тестов. Так как факторная нагрузка представляет собой еще и корреляцию между тестом и фактором, квадрат факторной нагрузки указывает нам долю общей дисперсии между тестом и соответствующим фактором. Приведенные в правой части табл. 11–3 квадраты факторных нагрузок показывают пропорциональный вклад каждого фактора в полную дисперсию показателей теста. Так, в тесте на арифметическое рассуждение 16 % дисперсии можно приписать вербальному, 30 % — числовому и 36 % —

[1] Это справедливо только для случаев, когда применяется ортогональное вращение. При облическом (косоугольном) вращении, речь о котором пойдет в этом разделе позднее, факторные нагрузки и факторные корреляции связаны между собой простым отношением, позволяющим с помощью соответствующих вычислений легко получить одно из другого.

Таблица 11–3

Источники дисперсии тестовых показателей

Тест	Нагрузки общего фактора			Коэффициент надежности	Относительный вклад				
	V	**N**	**R**		**V**	**N**	**R**	Специфический	Ошибка
1. Арифметическое рассуждение	0,40	0,55	0,60	0,90	0,16	0,30	0,36	0,08	0,10
2. Умножение	0,10	0,70	0,30	0,85	0,01	0,49	0,09	0,26	0,15

логическому факторам. Дисперсия ошибок в последнем столбце определена простым вычитанием коэффициента надежности из полной дисперсии (1,00 – 0,90 = 0,10). Цифры слева, указанные рядом с дисперсиями ошибок, отражают специфичность каждого теста, т. е. долю его «истинной» дисперсии, не разделяемую ни с одним другим тестом, вместе с которыми данный тест подвергался факторному анализу. Для теста на арифметическое рассуждение мы получаем следующие величины:

$$0,16 + 0,30 + 0,36 + 0,10 = 0,92$$
$$1,00 - 0,92 = 0,08$$

На рис. 11–2 структура полной дисперсии для двух тестов (в соответствии с данными табл. 11–3) представлена в графической форме.

Любой индивидуальный результат по этим двум тестам определяется величиной соответствующих способностей, или факторов, которыми обладает конкретный человек, а также относительными весами этих факторов в конкретном тесте. Поэтому если бы мы располагали чьими-то показателями по вербальному, числовому и логическому факторам, выраженными в одинаковых единицах измерения, то могли бы взвесить каждый показатель, умножая его на соответствующую факторную нагрузку. Сумма этих произведений дала бы нам оценку показателя данного человека по данному тесту. Чем меньше вклад специфического и случайного факторов в этот тест, тем точней будет наша оценка.

Согласно гипотетическим данным табл. 11–3, если конкретный человек имеет очень высокую оценку по Вербальному фактору (*V*), это гораздо больше поможет ему при выполнении теста на арифметическое рассуждение, чем теста на умножение. Фактически, содействие фактора *V* оказалось бы в 4 раза сильнее в тесте на арифметическое рассуждение по сравнению с тестом на умножение, поскольку вес этого фактора в 4 раза больше в первом тесте, чем во втором (0,40 против 0,10). Из трех общих факторов Числовой фактор (*N*) имел бы наибольшее влияние в тесте на умножение (нагрузка = 0,70), а Логический фактор (*R*) — в тесте на арифметическое рассуждение (нагрузка = 0,60).

Факторные нагрузки и корреляция. Вторая основная теорема ФА касается соотношения факторных нагрузок и корреляций между переменными. Корреляция между любыми двумя переменными равняется сумме парных произведений их нагрузок по

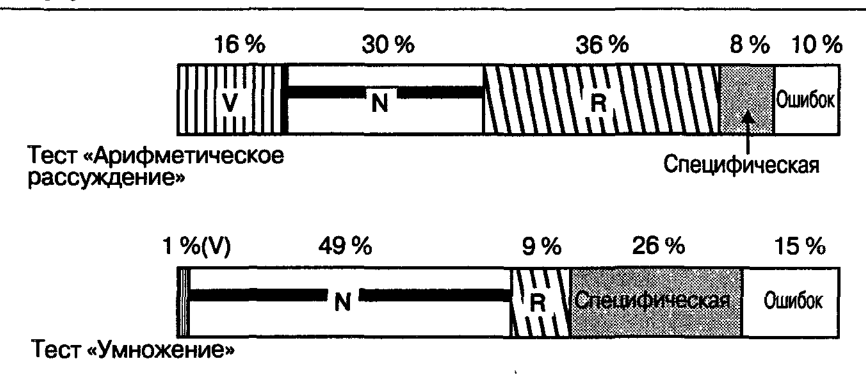


Рис. 11–2. Процентное соотношение общей дисперсии, специфической дисперсии и дисперсии ошибок в двух гипотетических тестах

(По данным табл. 11–3)

общим факторам. Так как специфический фактор и фактор ошибки каждой переменной уникальны, они не вносят никакого вклада в корреляцию между переменными. Корреляция между любыми двумя переменными зависит только от тех факторов, которые являются общими для этих двух переменных. Чем больше веса таких общих факторов в обеих переменных, тем выше будет между ними корреляция. Корреляцию между двумя тестами из табл. 11–3 можно найти перемножением нагрузок каждого из трех общих факторов по этим двум тестам и сложением полученных произведений:

$$r_{12} = (0{,}40)(0{,}10) + (0{,}55)(0{,}70) + (0{,}60)(0{,}30) = 0{,}60.$$

Косоугольная система координат и факторы второго порядка. Изображенные на рис. 11–1 оси называются *ортогональными,* так как они строго перпендикулярны друг другу. Иногда кластеры тестов располагаются таким образом, что лучшего соответствия используемым критериям удается достичь при использовании *облических, или косоугольных,* осей. В таком случае уже сами факторы коррелируют друг с другом. Одни исследователи утверждали, что использование ортогональных, или некоррелирующих, факторов всегда предпочтительнее, поскольку такие факторы дают более простую и четкую картину взаимосвязи черт. Другие настаивают на том, что косоугольную систему координат следует использовать всякий раз, когда она лучше соответствует изучаемым данным, поскольку большинство имеющих ясный физический смысл категорий и не должны быть независимыми. Очевидный пример — рост и вес. Хотя хорошо известно, что рост и вес высоко коррелируют между собой, они оказались весьма полезными категориями при оценке телосложения.

Когда факторы коррелируют между собой, существующие между ними интеркорреляции можно подвергнуть тому же статистическому анализу, который мы применяем к интеркорреляциям между тестами. Иными словами, у нас есть возможность «факторизовать факторы» и получить *факторы второго порядка.* Этот способ обработки данных был использован в ряде исследований таких переменных, как способности и черты личности. В некоторых исследованиях с использованием тестов способностей был получен единственный общий фактор второго порядка. Как правило, американские исследователи, применяющие факторный анализ, начинают с объяснения как можно большей части общей дисперсии групповыми факторами и только затем выявляют

общий фактор как фактор второго порядка, если данные подтверждают его наличие. У английских психологов, напротив, принято начинать с общего фактора, которому приписывается основная доля общей дисперсии, а затем возвращаться к групповым факторам для объяснения остаточной корреляции. Эта разница в методиках есть следствие теоретических различий, о которых речь пойдет в следующем разделе.

Теории организации черт

На протяжении более полувека предпринимались многочисленные попытки с помощью статистических методов ФА понять природу и организацию способностей, связанных с разнообразной человеческой деятельностью. Тем не менее эти методы до сих пор остаются наиболее тесно связанными с изучением когнитивных способностей, или «интеллекта», — направлением, в рамках которого и зародился факторный анализ. Недавно составленный обзор всех опубликованных факторно-аналитических исследований когнитивных способностей дает впечатляющую сводку состояния дел в этой области (Carroll, 1993). Охватывая 70-летний период исследований, работа Кэрролла представляет собой гораздо больше, чем литературный обзор, ибо содержит еще и повторный анализ 450 наборов данных из оригинальных исследований. К тому же в ней описаны различные теоретические модели интеллекта и дана их оценка в исторической перспективе. В этом разделе мы рассмотрим лишь некоторые широко известные теории интеллекта, выбор которых обусловлен их воздействием на конструирование и использование тестов.

Двухфакторная теория. Первой теорией организации черт, основанной на статистическом анализе показателей тестов, была двухфакторная теория, развитая английским психологом Чарльзом Спирменом (Spearman, 1904; 1927). В своем первоначальном виде эта теория утверждала, что все виды интеллектуальной активности используют долю единого общего фактора, названного *генеральным,* или фактором *g* (от англ. *general*). Кроме того, в теории Спирмена постулировалось наличие многочисленных *специфических*, или *s*-факторов (от англ. *specific*), каждый из которых сказывается на выполнении только одной из интеллектуальных функций. Положительная корреляция между любыми двумя функциями приписывалась, таким образом, действию фактора *g*. Чем больше эти две функции были «насыщены» (*saturated*) фактором *g*, тем выше должна бы быть корреляция между ними. Напротив, присутствие специфических факторов вело к снижению корреляции между функциями.

Несмотря на постулирование Спирменом двух типов факторов — генерального и специфических, фактор *g* рассматривается в его теории как единственная причина корреляции. Поэтому, в отличие от других теорий связи черт, эту теорию было бы точнее называть однофакторной, однако она сохранила свое первоначальное название. Рис. 11–3 иллюстрирует основополагающий принцип корреляций тестов согласно этой теории. Из этой схемы видно, что, в соответствии с теорией Спирмена, тесты 1 и 2 должны высоко коррелировать между собой, поскольку каждый сильно насыщен фактором *g*, о чем свидетельствуют заштрихованные участки. Незаштрихованным частям каждого теста соответствуют специфический фактор и дисперсия ошибок. Тест 3 должен слабо коррелировать с каждым из двух других тестов, поскольку включает очень малую долю фактора *g*.

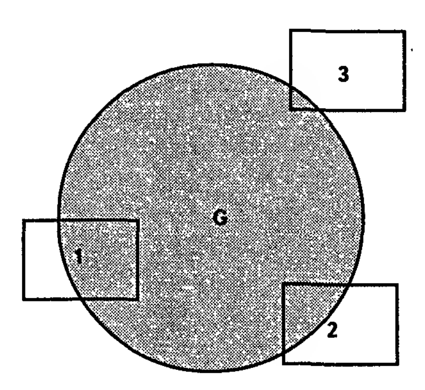


Рис. 11–3. Принципиальная модель корреляции в двухфакторной теории

Согласно двухфакторной теории, целью психологического тестирования должно быть измерение величины фактора *g* у каждого индивидуума. Если этот фактор пронизывает все способности, тогда он дает нам единственную основу для предсказания результатов деятельности индивидуума в разных ситуациях. Специфические факторы измерять бесполезно, так как каждый из них, по определению, сказывается только на какой-то одной функции. Вот почему Ч. Спирмен предложил заменить разнородную совокупность заданий, встречаемых в тестах интеллекта, единственным, пусть односторонним, тестом, но при этом высоко насыщенным фактором g. Он полагал, что тесты на абстрактные отношения, по всей вероятности, лучше всех других измеряют *g* и потому могут быть использованы для этой цели. Примерами тестов, разработанных для измерения *g*, являются Прогрессивные матрицы Равена и Культурно-свободный тест интеллекта Кэттелла (*Cattell's Culture Fair Intelligence Test*).

С самого начала Спирмен понимал, что двухфакторная теория нуждается в уточнении. Когда сравниваемые деятельности достаточно похожи, корреляция между ними может достигать величины, превышающей степень связи между переменными, объяснимую действием фактора *g*. Поэтому в добавление к генеральному и специфическим факторам, вероятно, существует промежуточный класс факторов, не столь универсальных, как *g*, но и не столь специфичных, как *s*-факторы. Такой фактор, общий только для группы (а не для всех вообще) интеллектуальных функций, был назван *групповым фактором.* В первых вариантах своей теории Спирмен допускал возможность весьма узких и пренебрежимо малых групповых факторов. Позднее, под давлением фактов, полученных в исследованиях некоторых его учеников, он стал использовать в своих теоретических построениях гораздо более широкие групповые факторы, такие как арифметические, технические и лингвистические способности.

Многофакторные теории. Преобладавший в американской психологии взгляд на организацию черт, основанный на ранних факторно-аналитических исследованиях, заключался в признании ряда довольно широких групповых факторов, каждый из которых мог входить с разными весами в различные тесты. Например, вербальный фактор мог бы иметь значительный вес в словарном тесте, несколько меньший вес — в тесте словесных аналогий, и еще меньший — в тесте на арифметическое рассуждение. На рис. 11–4 в наглядной форме представлены интеркорреляции пяти тестов с точки

зрения многофакторной модели. Корреляции тестов 1, 2 и 3 друг с другом — следствие их общих нагрузок вербальным фактором (*V*). Аналогично этому, корреляция между тестами 3 и 5 — результат действия пространственного фактора (*S*), а между тестами 4 и 5 — числового (*N*). Тесты 3 и 5 отличаются сложной факторной композицией: каждый имеет существенные нагрузки более чем по одному фактору (тест 3 — по факторам *V* и *S*, а тест 5 — по факторам *S* и *N*). Обращаясь к рассмотренной в предыдущем разделе второй основной теореме ФА, мы можем сделать некоторые выводы об относительной величине этих интеркорреляций. Например, тест 3 будет сильнее коррелировать с тестом 5, чем с тестом 2, потому что веса фактора *S* в тестах 3 и 5 (области с диагональной штриховкой) больше, чем веса фактора *V* в тестах 2 и 3 (области с горизонтальной штриховкой).

Публикация программной книги Т. Келли *Crossroads in the Mind of Man* (T. L. Kelly, 1928) подготовила почву для большого числа исследований, нацеленных на выявление групповых факторов. Важнейшими среди предложенного Келли набора факторов были следующие: манипулирование пространственными отношениями, легкость оперирования числами, легкость оперирования словесным материалом, а также память и скорость. Этот перечень был позднее переработан и дополнен исследователями, использовавшими более современные методы ФА, рассмотренные в предыдущем разделе.

Одним из ведущих представителей многофакторной теории был Л. Л. Тёрстоун. Основываясь на обширных исследованиях, как своих собственных, так и учеников, Тёрстоун выделил около дюжины групповых факторов, которые он назвал «первичными умственными способностями». К факторам, чаще всего подтверждавшимся в работах самого Тёрстоуна и других независимых исследователей (French, 1951; Harman, 1975; Thurstone, 1938; Thurstone, & Thurstone, 1941), относятся следующие:

V. Вербальное понимание (Verbal Comprehension). Главный фактор в таких тестах, как понимание прочитанного, словесные аналогии, восстановление порядка слов в предложениях, вербальное рассуждение и подбор пословиц. Данный фактор наиболее адекватно измеряется словарными тестами.

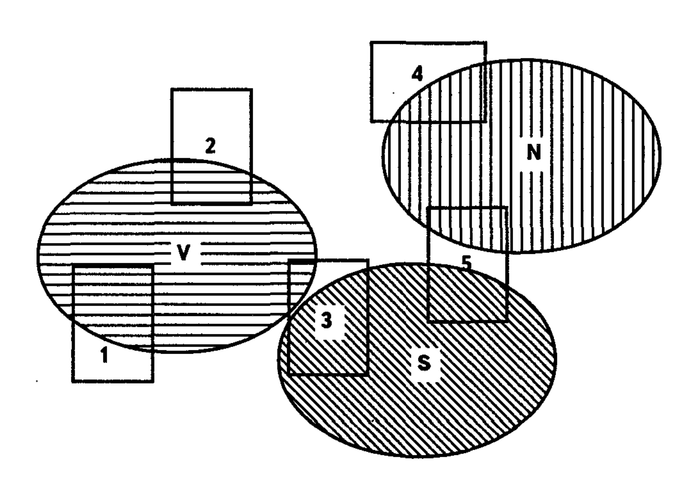


Рис. 11-4. Принципиальная модель корреляции в многофакторных теориях

W. Беглость речи (Word Fluency). Выявляется в таких тестах, как анаграммы, подбор рифм или называние слов данной категории (например, мужские имена или слова, начинающиеся с буквы Т).

N. Числовой (Number). Почти полностью отождествляется со скоростью и точностью простых арифметических вычислений.

S. Пространственный (Space). Может представлять собой два разных фактора. Один связан с восприятием фиксированных пространственных или геометрических отношений, другой с манипулированием зрительными образами, при котором изменение положения или трансформацию объекта необходимо представить зрительно (McGee, 1979; Portegal, 1982).

М. Ассоциативная память (Associative Memory). В основном обнаруживается в тестах, требующих механической памяти на ассоциативные пары. Есть некоторые основания предполагать, что этот фактор может отражать степень использования опор памяти (Christal, 1958), а противоречит существованию более широкого фактора, присутствующего во всех тестах памяти. Некоторые исследования подтверждают наличие ограниченных факторов памяти, таких как память на временные последовательности и положение в пространстве.

Р. Перцептивная скорость (Perceptual Speed). Быстрое и точное зрительное восприятие деталей, сходства и различий. Возможно, это тот же фактор, что и фактор скорости, выявленный Т. Л. Келли и другими предшественниками, по крайней мере, он относится к ряду факторов, идентифицированных позднее в задачах на восприятие (Thurstone, 1944).

I (или R). Индукция (или Общий вывод) — (Induction, or General Reasoning). Этот фактор установлен наименее четко. Тёрстоун первоначально предположил наличие индуктивного и дедуктивного факторов. Последний лучше всего измерялся тестами на силлогистический вывод, а первый — тестами, требующими от испытуемого найти принцип (правило, закономерность и т. п.), как в тестах на завершение числовых последовательностей. Доказательства наличия дедуктивного фактора оказались, однако, гораздо слабее доказательств в пользу существования индуктивного фактора. Кроме того, некоторые исследователи исходили из предположения, что фактор логического мышления лучше всего измеряется тестами на арифметическое мышление.

Следует отметить, что различия между общими, групповыми и специфическими факторами не столь существенны, как может показаться в первый момент. Если число или разнообразие тестов в батарее невелико, одним общим фактором можно объяснить все корреляции между ними. Но когда те же самые тесты включены в батарею с более разнородным составом тестов, исходный общий фактор может выделиться как групповой, т. е. общий только для некоторых, но не для всех тестов. Аналогично этому, некоторый фактор может быть представлен только одним тестом в исходной батарее, но разделяться несколькими тестами в более крупной батарее. Такой фактор был бы идентифицирован как специфический в первой батарее, но оказался бы групповым в более полной, комплексной батарее. Поэтому вряд ли нужно удивляться, что интенсивные факторные исследования специальных областей выявили множество факторов вместо одной или двух первичных умственных способностей, первоначально идентифицированных в каждой такой области. Именно это и произошло в исследованиях вербальных и перцептивных тестов, тестов памяти и тестов на логическое рассуждение.

Складывается представление, что факторные исследования привели к ошеломительному «размножению» факторов. Число когнитивных факторов, описанных на се-

годняшний день различными исследователями, перевалило за 100. Относительного порядка в этой сфере удалось достичь путем перекрестной идентификации факторов, описанных разными исследователями и зачастую под разными названиями (Ekstrom, French, & Harman, 1979; French, 1951; Harman, 1975). Такую перекрестную идентификацию можно выполнить только в тех случаях, когда в сравниваемых исследованиях используется ряд общих тестов. Чтобы облегчить этот процесс, группой сторонников применения факторного анализа был составлен комплект «базовых тестов» (*reference tests*), измеряющих установленные к настоящему времени главные факторы способностей. Этот комплект, распространяемый Службой тестирования в образовании (Ekstrom, French, Harman, & Dermen, 1976; ETS kit, 1976), облегчает разным исследователям планировать факторные исследования с включением ряда общих тестов в используемые ими батареи.

Очевидно, что даже после всех этих попыток упростить ситуацию и согласовать действия исследователей в области изучения способностей методами ФА, число факторов остается большим. Человеческое поведение изменчиво и сложно, и, по-видимому, наивно ожидать, что его можно адекватно описать с помощью дюжины или около того факторов. Но для конкретных целей можно подобрать подходящие факторы в отношении как их природы, так и их широты. Так, если бы мы отбирали кандидатов для трудной и высокоспециализированной работы технического характера, то, вероятно, захотели бы измерить у них довольно узкие факторы восприятия и пространственных отношений, наиболее отвечающие требованиям будущей работы. С другой стороны, при отборе студентов колледжа, мы бы отдали явное предпочтение нескольким широким факторам, таким как вербальное понимание, легкость оперирования числами и умение делать общие выводы.

Модель структуры интеллекта. Некоторые исследователи, работающие с факторным анализом, пытались упростить картину взаимосвязей черт путем придания им некой упорядоченной структуры. Основываясь на своих более чем 20-летних факторно-аналитических исследованиях, Дж. Гилфорд (Guilford, 1967, 1988; Guilford, & Hoepfner, 1971) предложил кубическую модель, которую назвал моделью «структуры интеллекта» (*structure-of-intellect model* или, сокращенно, *SI model*). Изображенная на рис. 11–5 модель[1] классифицирует черты интеллекта (*intellectual traits*) по трем измерениям.

Операции — то, что делает респондент. К ним относятся восприятие (*cognition*), запоминание и сохранение информации, дивергентное (проявляющееся в творческой активности) продуцирование, конвергентное продуцирование и оценка.

Содержание — характер материалов или информации, с которыми выполняются операции. Сфера содержания делится на зрительную, слуховую, символическую (например, буквы, цифры), семантическую (например, слова) и поведенческую (информация о других людях, их поведении, аттитюдах, потребностях и т. д.) области.

Продукты — форма, в которой информация обрабатывается респондентом. Продукты подразделяются на элементы, классы, отношения, системы, преобразования и импликации.

---

[1] В результате последующих исследований Гилфорда его модель претерпела изменения и отличается от более ранней (Guilford, 1967) разделением такого аспекта содержания, как образная (*figural*) информация, на зрительную и слуховую составляющие, а также разделением такого аспекта операций, как память, на запоминание и сохранение материала (Guilford, 1988).

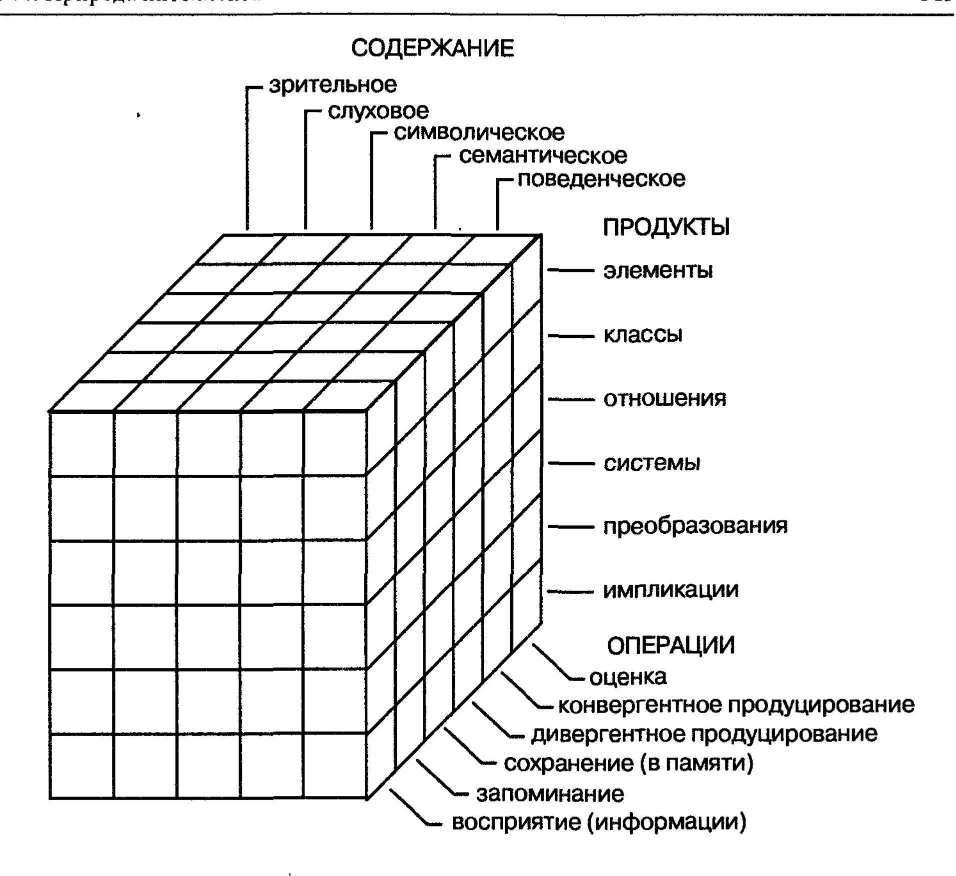


Рис. 11-5. Трехмерная модель структуры интеллекта
(Из Guilford, 1988, p. 3. Copyright © 1988 by Educational and Psychological Measurement. *Воспроизведено с разрешения)*

Поскольку такая классификация содержит 6 × 5 × 6 категорий, в этой модели получается 180 ячеек. Предполагается, что каждой ячейке соответствует по меньшей мере один фактор, или способность, а некоторые ячейки могут содержать более одного фактора. Каждый фактор описывается в трех измерениях. К моменту завершения 20-летнего Проекта исследования способностей (*Aptitude Research Project*) — скоординированной программы исследований *SI-model* – Дж. Гилфорд и его коллеги идентифицировали 98 из предсказанных на основе этой модели факторов (Guilford, & Hoepfner, 1971). Алфавитный перечень и краткое описание множества тестов, разработанных в рамках этого 20-летнего проекта, можно найти в книге Дж. Гилфорда и Р. Хопфнера (Guilford, & Hoepfner, 1971, Appendix B).

Хотя один тест все же был разработан непосредственно на основе *SI-model*, а именно — СИ-тест способностей к обучению (*Structure of Intellect Learning Abilities Test* — Meeker, Meeker, & Roid, 1985), сама эта модель практически не повлияла на разработку и применение тестов общего пользования.[1] Следует, однако, учитывать то обстоятельство, что *SI-model*, как и все другие модели организации черт, дает всего одну

[1] Подробнее об этом см. Carroll (1993), особенно р. 57–60.

схему отображения полученных корреляций между переменными. Вследствие метода, используемого при вращении осей, эмпирическое подтверждение *SI-model* не служит основанием для опровержения других моделей (Carroll, 1972; Horn & Knapp, 1973). Вполне вероятно, что различные вращения одних и тех же данных могут обнаружить одинаково тесное их соответствие другим моделям. Повторный анализ исходных факторно-аналитических данных Гилфорда действительно показал, что они лучше соответствовали другим моделям, чем его оригинальной *SI-model*, и при этом получали менее противоречивое теоретическое и практическое истолкование (Bachelor, 1989; Carroll, 1993).

С другой стороны, благоприятным косвенным влиянием проекта Гилфорда стало привлечение внимания к разграничению операций и содержания при идентификации факторов. Это разграничение помогло прояснить и сущность факторов, выделяемых с помощью ФА, и природу процессов, исследуемых в когнитивной психологии, и, что немаловажно, отношение между факторами и процессами. Несколько больше об этом будет сказано в последнем разделе этой главы, посвященном природе и развитию черт. Еще одним полезным побочным результатом разработки *SI-model* является различение конвергентного и дивергентного мышления. Последнее понятие, подразумевающее атипичное поведение, было широко принято на вооружение в анализе креативности. Однако попытки разработать тесты дивергентного мышления, независимые от сферы содержания, в целом оказались безуспешными. Как выяснилось, дивергентное мышление и творческая продуктивность зависят от того, в какой области или с каким содержанием работает человек, например, в области конкретной науки (скажем, биологии или физики) или в конкретной сфере искусства (музыки, скульптуры и т. д.). Более того, творческая идея или иной продукт творчества должны обрести смысл или принести пользу в рамках опять же конкретной культуры; сама по себе дивергенция, без качественного прогресса, не имеет к творчеству никакого отношения. Исследования широкой проблемы креативности, проявляемой в любых контекстах и формах, развивались быстрыми темпами. Рост интереса к этой области исследований иллюстрируется серией томов по вопросам креативности, опубликованных в 1990-х гг. под редакцией Р. С. Алберта (R. S. Albert, 1991–1994). Отдельные тома этой серии дают широкое освещение проблемы, от разнообразных проявлений креативности до способов стимуляции и развития креативного поведения.

Иерархические теории. Альтернативная схема организации факторов была предложена рядом английских психологов, включая С. Берта (Burt, 1949) и Ф. Вернона (Vernon, 1960), и американцем Л. Хамфрейсом (Humphreys, 1962). Схема, поясняющая применение Верноном этого подхода, воспроизведена на рис. 11–6. На вершине иерархии Вернон поместил спирменовский фактор *g*. На следующем уровне — два широких групповых фактора, соответствующих вербально-образовательным *(v : ed)* и практико-техническим способностям *(p : m)*. Эти главные факторы можно далее подразделить на несколько второстепенных. Вербально-образовательный фактор, например, дает среди прочих вербальный и числовой субфакторы, а практико-технический разделяется на такие субфакторы, как технической осведомленности, пространственный и психомоторных способностей. Еще более узкие субфакторы можно выделить в ходе последующего анализа, скажем, вербальных заданий. На самом нижнем уровне иерархии находятся специфические факторы. В более поздний, уточненный вариант этой модели Вернон (Vernon, 1969) включил более сложные взаимосвязи и перекрест-

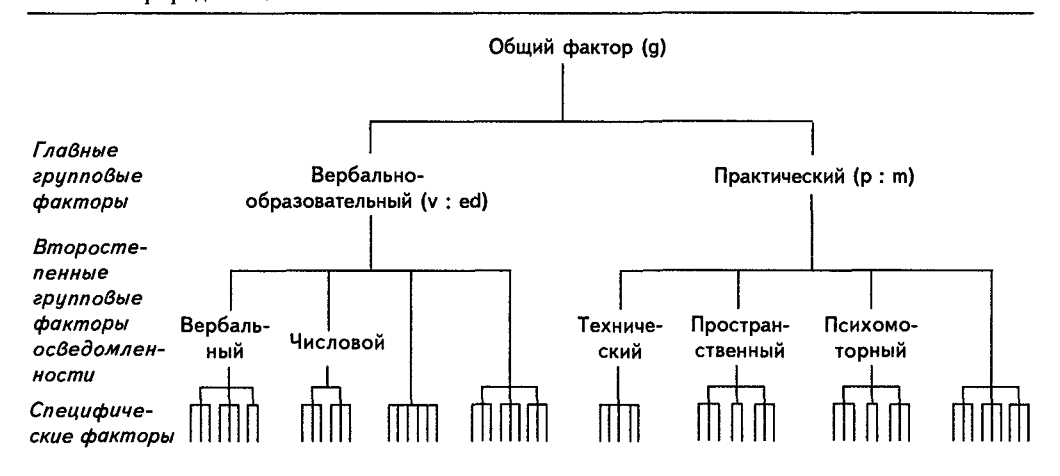


Рис. 11–6. Модель иерархической организации способностей
(С упрощениями из P. E. Vernon, 1960, p. 22. Copyright © 1960, Methuen & Co., Ltd. Воспроизводится с разрешения)

ные вклады факторов на третьем уровне, особенно в том, что касается образовательных и профессиональных достижений. К примеру, научные и технические способности связаны в этой модели с пространственными способностями и технической осведомленностью; математические способности — с пространственными и числовыми, а также, почти напрямую, с фактором *g* (через фактор индукции).

Л. Хамфрейс (Humphreys, 1962, 1970) также рекомендовал иерархическую модель в качестве средства, позволяющего справиться с разрастанием факторов. Однако, вместо того чтобы считать какой-то один уровень факторов главным (или первичным), он предлагал составителям или пользователям тестов выбирать тот уровень иерархии, который наиболее соответствует их целям. Кроме того, Хамфрейс признавал, что один и тот же тест, в зависимости от содержания, процесса и других аспектов, может быть внесен более чем в одну иерархию. По его мнению, чтобы измерить какой-нибудь один аспект, нужно сделать тест *гетерогенным* относительно всех остальных аспектов. Если, например, нас интересует способность человека решать задачи на аналогии, то следует воспользоваться тестом, содержащим вербальные, числовые, рисуночные и пространственные аналогии. Если же мы хотим измерить вербальную способность, нам следует использовать разнообразные типы заданий, такие как определение слов, аналогии и завершение рядов. Эта методика отличается от той, которой пользовался Гилфорд, искавший отдельные факторы (и тесты) для каждой гомогенной ячейки своей трехмерной классификации. Однако в своей более поздней работе Гилфорд (Guilford, 1981) применил схему частичной иерархической организации при идентификации факторов высшего порядка среди некоторых факторов, входящих в его оригинальную модель структуры интеллекта.

Иерархическая модель интеллекта получает все более широкое признание как по теоретическим, так и по практическим соображениям (см. Anastasi, 1992a, 1994; Carroll, 1993; Gustafsson, 1984, 1989; Lubinski, & Dawis, 1992). Как теоретическая модель связи черт она совмещает единственный общий фактор (*g* Спирмена) с многофакторными отображениями. В методологическом плане было доказано, что многофактор-

ные и иерархические решения математически эквивалентны и допускают преобразование одного в другое (Harman, 1976; chap. 15; Schmid, & Leiman, 1957). Косоугольное решение (с коррелируемыми факторами), которое приводит к иерархической модели, можно преобразовать в ортогональное решение (с некоррелируемыми факторами). В ортогональном решении факторы второго порядка выделяются как факторы первого порядка иной широты. Более широкие факторы имеют нагрузки по большему числу переменных, чем менее широкие.

С практической точки зрения, главное преимущество тестов, разработанных исходя из иерархической модели, состоит в том, что они сочетают всесторонний охват способностей с гибкостью использования. Сообразуясь с различными целями тестирования, пользователь может выбрать один суммарный показатель батареи либо один или несколько показателей по кластерам тестов, измеряющим более узко определяемые факторы. При определенных обстоятельствах показатели по отдельным субтестам также могут оказаться полезными, например, в том, что касается выявления слабости или силы специализированных навыков.

Возрастающее влияние иерархической модели на конструирование и использование тестов способностей отмечалось в главе 2 как линия исторического развития тестирования, ведущая к сближению традиционных тестов интеллекта с комплексными батареями способностей. Примеры иерархического подхода при разработке индивидуальных тестов приведены в главе 8, групповых — в главе 10. К наиболее ясным примерам такого рода относятся Дифференциальные шкалы способностей (Elliott, 1990b) для индивидуального тестирования и Многоаспектная батарея способностей (Jackson, 1994b) для группового тестирования. Заслуживающей упоминания мерой обеспечения гибкости в выборе уровней подсчета показателей, предусмотренной в таких тестах, является предоставление норм для интерпретации показателей на выбранном уровне. Без такого рода разнесенных по уровням норм правильная оценка выполнения теста конкретным человеком на разных уровнях иерархии была бы невозможна.

Природа и развитие черт

То, что разные исследователи могу приходить в процессе работы к столь непохожим моделям организации черт, не так обескураживает, когда мы осознаем, что выявляемые с помощью факторного анализа черты — это не более чем выражение корреляций между мерами поведения. К ним следует относиться не как к первоэлементам или причинным факторам, а как к описательным категориям. Отсюда понятно, что различные принципы классификации могут применяться к одному и тому же набору данных. Понятие факторов как описательных категорий было подробно разработано в ранних публикациях англичан Г. Томсона (Thomson, 1948), С. Берта (Burt, 1941, 1944), Ф. Вернона (Vernon, 1960) и американца Р. Трайона (R. C. Tryon, 1935). Все эти авторы обращали внимание на широкое разнообразие элементов поведения, которые могут образовывать группы («кластеры») под влиянием наследственности или благодаря приобретенным связям.

Разнообразие жизненного опыта. Имеет место растущее признание роли жизненного опыта конкретного человека в развитии интеллекта и формировании групповых

факторов (Anastasi, 1986b; Greeno, 1989). От того, как складывается жизнь, зависит не только уровень реализации различных способностей, но и способ организации характеристик деятельности в различные черты. Различия в факторных отображениях связывают с различными культурами или субкультурами, социоэкономическими уровнями или типами школьных программ (см. Anastasi, 1970, 1983a, 1986b, 1994; Vernon, 1969). Существенные изменения в факторных отображениях происходят и с течением времени. К ним относятся как долговременные, отражающие кумулятивное влияние опыта повседневной жизни, так и кратковременные, возникающие в результате практики и других экспериментально контролируемых случаев научения (Baltes, Cornelius, Spiro, Nesselroade, & Willis, 1980; Birren, Cunningham, & Yamamoto, 1983; Fleishman, 1972; Fleishman, & Mumford, 1989; Khan, 1970, 1972; Reinert, 1970). Исследования на животных также подтвердили возможность экспериментального получения факторов при контролировании раннего опыта (Whimbey & Denenberg, 1966).

Факторная композиция одной и той же объективной задачи может различаться у лиц с разным жизненным опытом. Одно из объяснений этих индивидуальных различий может быть найдено в использовании разных методов выполнения одной задачи. Лица с высокоразвитыми вербальными способностями, например, склонны использовать вербальные вспомогательные средства (*mediators*) для решения механических или пространственных задач; в свою очередь, те, чей опыт был связан с механикой, будут пытаться решить те же задачи, опираясь на восприятие или пространственные представления. Соответствующие доказательства были получены Френчем (French, 1965), установившим, что факторная композиция одних и тех же тестов различалась у групп лиц, объединенных в соответствии с их типичным стилем решения задач. Подтверждающие данные можно также найти в исследовании Фредериксена (C. H. Frederiksen, 1969), изучавшего когнитивные стратегии, используемые испытуемыми при запоминании слов. В процессе научения люди могут изменять свой выбор стратегии, что приводит к изменению факторной композиции выполняемой ими задачи. Доказано, что аналогичные изменения в обращении к специальным способностям происходят на протяжении более длительных периодов времени, в ходе обучения (R. B. Burns, 1980).

Механизмы формирования черт. Механизм возникновения факторов получает объяснение в таких известных понятиях, как установка на научение (*learning set*) и перенос навыков (*transfer of training*) (Carroll, 1966; G. A. Ferguson, 1954, 1956; Simon, 1990; Whiteman, 1964). Образование установок облегчает научение, когда предъявляется каждая новая задача того же типа. В классических экспериментах Г. Харлоу (Harlow, 1949, 1960) с обезьянами, после того как животное решало серию задач на различение определенных форм, скажем треугольника и круга, оно научалось различать *другие* формы гораздо быстрее, чем это делали животные, не имевшие такого предварительного опыта. У такого животного образовалась установка на научение различению форм, и оно уже «знало», чего ожидать при встрече с новой задачей. Таким образом, это животное «научилось учиться» решать задачи такого типа.

Точно так же большинство навыков, развитых во время обучения в школе, например чтение и арифметические вычисления, пригодны для использования в последующих достаточно разнообразных учебных ситуациях. Эффективные и систематические методы решения задач могут также применяться для решения новых задач. Индивидуальные различия в степени овладения этими навыками будут отражаться на качестве выполнения большого числа разнообразных задач, а при факторном анализе

этих задач такие широко применимые навыки могли бы проявиться в виде широких групповых факторов. Поэтому широтой переноса, или степенью разнообразия задач, к которым применим данный навык, и будет определяться широта получаемого группового фактора.

Другим важным источником формирования черт является сопряженность (*contiguity*) или со-распространенность (*co-occurrence*) ситуаций обучения. Например, у людей, принадлежащих к культурам с развитой системой образования, широкий вербально-образовательный фактор формируется, вероятно, в процессе овладения всем тем, чему их обучают в школе. Более узкий фактор числовых способностей может быть следствием того обстоятельства, что всем арифметическим действиям обучает один учитель и все это происходит в одном и том же классе. Поэтому ребенок, который падает духом, сопротивляется или откровенно скучает во время занятий арифметикой, вероятно, будет отставать в овладении *всеми* этими действиями, а тот, кто стимулирован и получает удовольствие от уроков арифметики, скорее всего хорошо усвоит все, что ему преподают на этих уроках, и разовьет способности, впоследствии дающие ему преимущество в овладении более сложными навыками оперирования числами.

Впрочем, каков бы ни был механизм их формирования, факторы, или способности, идентифицируемые с помощью факторного анализа, представляют собой дескриптивные категории, отражающие изменяющиеся взаимосвязи характеристик деятельности в разнообразных ситуациях. Эти факторы есть не застывшие сущности, а продукт накапливаемого человеком жизненного опыта. И коль скоро структура опыта варьирует у отдельных людей или их групп, разумно ожидать появления различных факторных отображений. По мере трансформации опыта конкретного человека — вследствие образования, выполнения профессиональных обязанностей или других продолжительных видов деятельности — могут появляться новые черты, а ранее существовавшие — сливаться в более широкие комплексы.

Факторный анализ и когнитивный анализ задачи. Применение методов анализа обработки информации, разработанных значительно позднее в рамках когнитивной психологии, вносит существенный вклад не только в понимание того, что измеряют тесты интеллекта (см. главу 5), но и позволяет лучше понять процесс формирования и развития факторов. Метод анализа протоколов, когда испытуемого побуждают «думать вслух» во время решения задачи или выполнения задания в уме, расценивается как многообещающий подход к изучению человеческого мышления (Ericsson & Simon, 1993). Однако в ходе продолжения исследований когнитивных процессов накапливается все больше доказательств *зависимости процессов мышления от предметной области*. За исключением самых элементарных уровней, навыки обработки информации все же специфичны для обрабатываемого типа содержания, причем приобретаются по мере овладения этим содержанием и его организации для быстрого поиска в памяти.

С другой стороны, наиболее часто выявляемые с помощью факторного анализа черты имеют отношение, главным образом, к областям содержания, таким как вербальная, числовая и пространственная. А менее общие факторы, которые определяются через процессы, такие как ассоциативная память, индукция или дивергентное мышление, сами собой оказываются зависящими от предметной области, когда оцениваются посредством специально разработанных тестов. Независимо от того, построен ли

тест на основе факторного или когнитивного анализа, основное различие — это различие между процессом и той областью, в которой он используется. Релевантная область может относиться к обрабатываемому *содержанию* (например, лингвистическому, математическому, механическому и т. д.) или к *контексту*, в котором осуществляется процесс обработки содержания, описываемому с помощью культурологических, социологических, географических, профессиографических и других категорий, предназначенных для характеристики окружающей среды.

Прогресс когнитивной психологии, от ранних попыток идентифицировать «абсолютные процессы» человеческого интеллекта до стремительно развивающихся поисковых исследований предметной специфичности всех когнитивных процессов, отражается во впечатляющем росте публикаций, освещающих специфичность когнитивных процессов в зависимости от предметной области.[1] Примечательно также, что тесты, разработанные для оценки содержательно определяемых черт (*content-defined traits*), установленных в результате факторного анализа, оказались хорошими предикторами повседневных дел. Вполне вероятно, что такие факторы представляют собой опосредованно идентифицируемые кластеры навыков обработки информации, соответствующих данной содержательной области. Поэтому человек, получивший высокий показатель по тесту вербальных способностей, может выделяться не только объемом и организацией вербальных знаний, но и использованием специализированных навыков обработки информации, без которых в вербальной области не обойтись.

Общий интеллект. Целью факторно-аналитических исследований интеллекта было не только выяснение того, что измеряют интеллектуальные тесты, но прояснение сущности интеллекта в любых его проявлениях. При рассмотрении с этой более широкой точки зрения, интеллект человека представляется этаким соединением когнитивных навыков и знаний, востребуемых, прививаемых и вознаграждаемых жизненной средой каждого конкретного человека (Anastasi, 1986c). В этом обобщенном конструкте интеллекта предметная специфичность (*domain specificity*) оказывается даже более сущностным свойством, чем в более узких процессах, выявляемых протокольным или факторным анализом. Многие стороны интеллекта развиваются в то время, как индивидуум обучается фактуальным знаниям и умениям обрабатывать информацию в какой-то одной предметной области, такой как конкретная культурная среда или определенная профессиональная область.

Первые попытки выявить и определить универсальный, пригодный на все случаи жизни, интеллект принесли нам традиционные «тесты интеллекта». Вскоре такие тесты стали называть мерами академического интеллекта или способности к обучению, поскольку они измеряли именно этот частный вид интеллекта. Затем был предложен термин «практический интеллект», под которым понимался совершенно иной вид интеллекта, не охватываемый традиционными тестами (см. Anastasi, 1986b; Neisser, 1976; Sternberg, & Wagner, 1986). Такой практический интеллект оказался, однако, не единым и неделимым, а множеством разновидностей интеллекта, применимых в различных практических областях (Lave, 1988; Rogoff, & Lave, 1984). Теперь уже никто не говорит о двух типах интеллекта, академическом и практическом, поскольку стало ясно, что интеллект — это многоаспектный конструкт.

[1] См., например, Greeno (1989), E. Hunt (1987), Schneider, & Weinert (1990), Simon (1990), Sternberg, & Frensch (1991).

Слабое место традиционного факторного анализа — в недостаточном внимании к выбору анализируемых переменных (Anastasi, 1988a). Очевидно, что выявляемые в процессе такого анализа факторы получаются из интеркорреляций между выбранными переменными. Большинство исследований интеллекта начиналось с получения набора показателей по тестам, предназначенным для измерения того, чему обучают в школе. Поэтому и получаемые факторы отображали академический интеллект. В культурах с развитой системой образования такие факторы оказываются хорошими предикторами школьных достижений; кроме того, они дают умеренно высокие корреляции с уровнем выполнения многих работ, для которых значимо школьное образование. Однако если мы хотим оценить интеллект в более широких контекстах, нам нужно начинать с характеристик деятельности людей, которой они занимаются в реальной жизни и которая ценится в конкретной культуре. Хотя учащиеся школ — это, пожалуй, самый доступный контингент для проведения тестирования, становится очевидным, что к таким исследованиям нужно привлекать взрослых, занятых в различных профессиях. Исследования можно организовать так, чтобы они были выгодны участникам и давали научные данные в хорошо контролируемых условиях. Можно только приветствовать увеличение числа примеров таких основанных на сотрудничестве с испытуемыми исследований, успешно проводимых в промышленности и сельском хозяйстве (например, Fleishman, & Reilly, 1992b; Lubinski, & Dawis, 1992; Whyte, 1991).

Ряд разработанных в когнитивной психологии методов сами по себе акцентировали предметную специфичность интеллекта. Один из них — исследование когнитивных процессов путем анализа протоколов контрастных групп, таких как «эксперты» и «новички» в определенной сфере деятельности (например, шахматные игроки, машинистки, программисты). Неоднократно подтверждалось, что уровень исполнения и использование когнитивных навыков у конкретного человека строго специфичны для определенной области деятельности. Особое преимущество этого метода когнитивной психологии заключается в том, что исследования с его использованием проводились на широком множестве проявлений интеллектуальной активности человека в реальной жизни, от запоминания заказа официантом до постановки медицинского диагноза и вынесения решений в суде (Chi et al., 1988; Ericsson, & Smith, 1991).

Понятие предметной специфичности (*domain specificity*) оказывает все большее влияние на психологическое изучение различных вопросов. Например, оно ясно заметно в современных определениях гениальности и одаренности. Критерий для выявления одаренных детей теперь сместился от строго заданного уровня *IQ* (Termen et al., 1925) к выдающимся способностям в любой из множества социально желательных областей (Csikszentmihalyi, Rathunde, & Whalen, 1993; Feldman, & Bratton, 1972; Horowitz, & O'Brien, 1985; Subotnik, & Arnold, 1994). Аналогично этому, первые попытки разработать общий способ обработки тестов креативности или дивергентного мышления сразу натолкнулись на проблему предметной специфичности (Baer, 1993; Runco, 1991, 1994; Subotnik, & Arnold, 1994). К настоящему времени уже достигнут некоторый прогресс в интегрировании предметных областей в тестах креативности. По-видимому, союз когнитивной психологии с психометрическим факторным анализом не только обогатил наше понимание интеллектуальной деятельности, но и приблизил оба вида исследований к реалиям повседневной жизни.

12 ПСИХОЛОГИЧЕСКИЕ ПРОБЛЕМЫ ТЕСТИРОВАНИЯ СПОСОБНОСТЕЙ

Неизбежным следствием расширения и усложнения любого научного предприятия является усиливающаяся специализация интересов и функций его участников. Такая специализация явно видна в отношении психологического тестирования к основному руслу (*mainstream* — основное направление, главная линия, господствующая тенденция, например в искусстве, литературе, моде и т. п.) современной психологии (A. Anastasi, 1967, 1971). Специалисты в области психометрии, разрабатывая методы конструирования тестов, достигли в этом деле поистине невиданных высот. Но, поставляя на научный рынок технически совершенные измерительные инструменты, они относительно мало заботились о том, чтобы сопроводить их еще и психологической информацией, необходимой для правильного пользования такими инструментами. В результате устаревшие интерпретации результатов теста слишком часто продолжают жить в умах пользователей безотносительно к уже имеющимся данным исследований в соответствующей области поведения. Эта частичная изолированность психологического тестирования от других областей психологии — и как следствие неправильное использование и интерпретация тестов — объясняют отчасти недовольство общественности по поводу психологического тестирования, появившееся в 1950-х гг., резко возросшее в 1970-х и сохранившееся в отношении его многих его приложений до сегодняшнего дня. Темы, выбранные для обсуждения в этой главе, позволяют проиллюстрировать, каким образом результаты психологических исследований могут содействовать эффективному использованию тестов способностей и помочь исправить распространенные ошибочные представления об *IQ* и других аналогичных показателях.

Лонгитюдные исследования интеллекта детей

Важный путь к пониманию конструкта «интеллект» — лонгитюдные исследования одних и тех же лиц на протяжении длительного периода времени. Хотя такие исследования можно рассматривать как способствующие долговременной прогностической валидности конкретных тестов, они позволяют также делать более общие выводы от-

носительно природы интеллекта и смысла показателей интеллектуальных тестов. Когда считалось, что интеллект в значительной степени является выражением наследственного потенциала, то на *IQ* смотрели как на нечто остающееся практически неизменным на протяжении всей жизни человека. Любые наблюдавшиеся при повторном тестировании вариации *IQ* приписывались недостаткам измерительного инструмента: либо недостаточной надежности, либо плохому отбору тестируемых функций. Однако по мере исследования природы интеллекта пришло понимание того, что интеллект как таковой является сложным и динамичным образованием. В последующих разделах будут рассмотрены типичные результаты лонгитюдных исследований интеллекта и проанализированы условия, способствующие как устойчивости, так и неустойчивости тестируемых способностей.

Устойчивость результатов тестирования интеллекта. Накоплено огромное количество данных, показывающих, что на протяжении периода обучения в начальной, средней и высшей школе выполнение интеллектуальных тестов одними и теми же людьми остается довольно устойчивым (см. Anastasi, 1958, p. 232–238; Bornstein & Krasnegor, 1989; McCall, Appelbaum, & Hogarty, 1973). Например, в одном из первых в Швеции исследований относительно случайной популяции Т. Хазен (Husén, 1951) получил корреляцию 0,72 между тестовыми показателями 613 мальчиков-третьеклассников и их же показателями, полученными через 10 лет при поступлении на службу в армию. Несколько позже еще один шведский исследователь К. Хёрнквист (K. Härnqvist, 1968) сообщил о корреляции 0,78 между тестами, проведенными в 13- и 18-летнем возрасте на более чем 4500 лицах мужского пола. Даже тестирование дошкольников дает на удивление высокие корреляции с более поздним повторным тестированием. В лонгитюдном исследовании 140 детей, проведенном в Исследовательском институте Фелса (*Fels Research Institute*), корреляция показателей по шкале Стэнфорд—Бине, полученных в возрасте 3 и 4 лет, составила 0,83 (Sontag, Baker, & Nelson, 1958). Корреляции показателей тестов 3-летних детей с их более поздними показателями уменьшались с увеличением временно́го интервала между тестированиями, но к 12 годам они все еще были достаточно высоки — 0,46. Что касается шкалы Стэнфорд-Бине, особое значение имеет дополнительное исследование, проведенное К. Бредвеем, К. Томпсоном и Р. Крейвенсом (Bradway, Thompson, & Cravens, 1958) на детях из выборки стандартизации Стэнфорд—Бине 1937 г., впервые тестировавшихся в возрасте от 2 до 5,5 лет. Корреляция исходных показателей (*IQ*) этих детей с результатами их повторного тестирования через 10 лет составила 0,65, через 25 лет — 0,59. Корреляция между ретестами через 10 лет (средний возраст 14 лет) и через 25 лет (средний возраст 29 лет) равнялась 0,85.

Как и следовало ожидать, ретестовые корреляции тем выше, чем короче временной интервал между тестами. Кроме того, при постоянном ретестовом интервале эти корреляции имеют тенденцию повышаться с увеличением возраста детей. Влияния возраста и ретестового интервала на величину ретестовых корреляций проявляются в виде устойчивой закономерности и сами по себе достаточно предсказуемы (R. L. Thorndike, 1933, 1940). Одно из объяснений увеличивающейся с возрастом устойчивости показателей тестов интеллекта связано с *кумулятивным характером интеллектуального развития*, — ведь в каждом возрасте интеллектуальные навыки и знания индивидуума включают все его прежние навыки и знания плюс некий прирост за счет новых приобретений. Даже если эти ежегодные приобретения никак не связаны между со-

бой, стабилизация уровня выполнения теста с возрастом могла бы происходить просто в силу того, что более ранние приобретения по мере взросления составляют все большую часть совокупного объема навыков и знаний индивидуума.

Хотя такое частичное перекрытие навыков и знаний в соседних возрастах можно рассматривать как одну из причин повышения устойчивости тестовых показателей у развивающегося индивидуума, стоит указать на два дополнительных условия. Первое — это *стабильность окружающей среды,* свойственная периоду развития большинства людей. Следовательно, любые (благоприятные или неблагоприятные) условия, характерные для какой-либо стадии возрастного развития ребенка, обычно сохраняются в интервале между первичным и вторичным проведением теста. Дети обычно растут в одной семье и в относительно устойчивой социоэкономической и культурной среде. Для них нетипичны случайные перемещения из среды, стимулирующей интеллектуальное развитие, в тормозящую его среду, и наоборот.

В то же время следует заметить, что психологическая среда братьев и/или сестер, растущих в одной семье, существенно различается. Постепенно накапливается все больше с трудом собираемых данных, показывающих различные аспекты, в которых может различаться субъективный опыт воспитываемых вместе сиблингов (Boer, & Dunn, 1992; Dunn, & Plomin, 1990; Hetherington, Reiss, & Plomin, 1993). Само по себе наличие младшего или старшего сиблинга — это уже совершенной иной психологический опыт по сравнению с тем, когда кому-то выпадает быть единственным ребенком в семье. Характер родительских приемов воспитания, равно как и реакции родителей на поведение ребенка, тоже могут существенно различаться для сиблингов, родившихся в разное время. Кроме того, важные события, затрагивающие жизнь семьи, — такие как развод, резкое изменение уровня доходов или переезд из сельской местности в крупный город, — могут по-разному влиять на детей просто потому, что они находятся на разных стадиях возрастного развития. Ко всем этим потенциальным различиям нужно добавить опыт, накопленный разновозрастными сиблингами за пределами семьи. Следовательно, хотя постоянство семейной среды и вносит свой вклад в устойчивость результатов тестирования конкретного ребенка, оно не обязательно должно вести к сходству между сиблингами.

Второе условие, способствующее общей устойчивости результатов выполнения теста интеллекта конкретным лицом, связано с ролью *учебных навыков как предпосылки* дальнейшего обучения. Индивидуум не только сохраняет ранее усвоенное, но многое из того, чему он научился, снабжает его средствами для дальнейшего обучения. Таким образом, чем больше ребенок преуспел в приобретении интеллектуальных навыков и знаний в любой точке возрастного развития, тем больше пользы сможет он извлечь из опыта последующего обучения. Понятие готовности к обучению служит выражением этого общего принципа. Последовательный характер научения подразумевается и в уже обсуждавшемся подходе Ж. Пиаже к умственному развитию, а также в различных индивидуализированных учебных программах.

Тот же принцип положен в основу проекта Head Start и других программ выравнивания для детей дошкольного возраста, воспитывавшихся в неблагоприятной образовательной среде (Stanley, 1972, 1973; Zigler, & Valentine, 1980). Поскольку у таких детей отсутствуют некоторые существенные предпосылки для эффективного учения в школе, по мере перехода из класса в класс они, как правило, будут все сильнее отставать в усвоении школьной программы. Следует добавить, что предпосылки эффективного учения включают не только такие интеллектуальные навыки, как владение

языком и количественными понятиями, но также аттитюды, интересы, мотивацию, стили решения проблем, реакции на фрустрацию, представление о себе (Я-концепцию) и другие качества личности. Цель выравнивающих образовательных программ — обеспечить предпосылки для эффективного учения, которые дадут детям возможность извлекать пользу из последующего обучения в школе. Разумеется, преследуя такую цель, эти программы предполагают подрыв «стабильности *IQ*», который в противном случае остался бы низким.

Неустойчивость результатов интеллектуальных тестов. Корреляционные исследования тестовых показателей снабжают нас актуарными данными, пригодными для групповых предсказаний. По указанным выше причинам, показатели тестов стремятся к устойчивости в актуарном смысле. Вместе с тем изучение отдельных лиц может обнаружить у них значительные колебания тестовых показателей. Резкое увеличение или снижение показателей может происходить в результате серьезных изменений среды, в которой живет ребенок. Коренные изменения в структуре семьи или домашней обстановке, помещение в детский дом, тяжелая или продолжительная болезнь, лечебные или исправительные программы — вот примеры событий, могущих изменить последующее интеллектуальное развитие ребенка. Тем не менее даже те дети, жизненная среда которых остается неизменной, при повторном тестировании могут демонстрировать значительное увеличение или снижение показателей. Как нетрудно понять, эти изменения означают, что конкретный ребенок развивается быстрее или медленнее, чем нормативная популяция, на которой данный тест был стандартизован. В общем, дети, растущие в неблагоприятной образовательной среде, обнаруживают тенденцию к снижению тестовых показателей с возрастом, в то время как у детей из благоприятной в образовательном отношении среды показатели тестов интеллекта имеют тенденцию к повышению. В современных исследованиях уделяется все больше внимания изучению специфических характеристик среды и самих детей (см., например, Carroll, 1993, р. 669–674; Detterman, & Sternberg, 1982).

Обширные данные о величине изменений индивидуальных показателей при неоднократном выполнении теста интеллекта были впервые получены в Калифорнийском исследовании (*California Guidance Study* — Honzik, Macfarlane, & Allen, 1948). При анализе результатов повторного тестирования 222 участников этого исследования обнаружились изменения величины *IQ*, доходящие у отдельных лиц до 50 баллов. В период от 6 до 18 лет, когда ретестовые корреляции обычно высоки, у 59 % детей *IQ* изменился на 15 и более, у 37 % — на 20 и более и у 9 % — на 30 и более баллов. Большинство этих изменений не носило случайного, беспорядочного характера. Напротив, дети на протяжении ряда лет обнаруживали устойчивую тенденцию к увеличению или снижению *IQ*, и эти изменения были связаны с характеристиками среды. В Калифорнийском исследовании подробное изучение домашней обстановки и взаимоотношений между родителями и детьми показало, что значительные изменения величины *IQ* были связаны с культурной средой и эмоциональным климатом, в которых воспитывался ребенок. Дополнительное исследование, проведенное с участниками Калифорнийского проекта, достигшими 30-летнего возраста, по-прежнему выявило значимые корреляции между тестовыми показателями и семейной атмосферой, оцененной еще на 21-м месяце их жизни (Honzik, 1967). Внимание и интерес родителей к успехам ребенка в обучении оказались столь же важными коррелятами уровня

выполнения интеллектуальных тестов в последующем, как и другие переменные, отражающие родительскую заботу об общем благополучии ребенка.

Некоторые исследователи уделяли повышенное внимание характеристикам личности, связываемым с ускорением и замедлением интеллектуального развития. В Исследовательском институте Фелса 140 детей были включены в интенсивное лонгитюдное исследование, охватывающее период жизни от младенчества до юности (Kagan, & Freeman, 1963; Kagan, Sontag, Baker, & Nelson, 1958; Sontag et al., 1958). В этой группе дети, показавшие самый большой прирост и спад *IQ* в период от 4,5 до 6 лет, сравнивались по широкому набору характеристик личности и жизненной среды; то же самое было проделано и в отношении детей с наибольшими изменениями *IQ* в период от 6 до 10 лет. В дошкольные годы эмоциональная зависимость (*emotional dependency*) от родителей была основным условием, связанным со снижением величины *IQ*. В школьные годы прирост *IQ*, был связан главным образом со стремлением к высоким достижениям, соревновательными мотивами и любознательностью (в отношении природы). Многообещающие данные были также получены в отношении роли способностей самих родителей и характерных для них методов воспитания в формировании у ребенка этих черт.

При последующем анализе той же выборки, проведенном по достижении участниками 17 лет, внимание исследователей было сосредоточено главным образом на характере (или «паттернах») изменений интеллектуальных показателей со временем (McCall et al., 1973). У детей с различающимися паттернами динамики тестовых показателей сравнивались методы их домашнего воспитания, оцениваемые на основе периодических посещений их семей. В целом, полученные данные говорят о том, что родители детей, обнаруживших тенденцию к росту интеллектуальных показателей, в дошкольные годы обычно «ободряли и поощряли своих детей, не забывая при этом устанавливать некоторые ограничения и добиваться их соблюдения» (McCall, 1973, р. 54). Главное условие, связываемое с положительной возрастной динамикой интеллектуальных показателей, описывается как попытка акселерации (*accelerational attempt*), или то, в какой степени «родители сознательно формировали у ребенка различные, пока еще не необходимые ему умственные и двигательные навыки» (р. 52).

Исследования факторов, связанных с ростом и снижением показателей тестов интеллекта, проливают свет на условия, определяющие интеллектуальное развитие в целом. Кроме того, они наводят на мысль, что предсказание последующего интеллектуального статуса можно улучшить, если исходные тестовые показатели сочетать с эмоциональными и мотивационными характеристиками индивидуума и с параметрами его жизненной среды. Согласно еще одной точке зрения, результаты такого типа исследований указывают путь к разработке конкретных программ вмешательства, способных эффективно влиять на ход интеллектуального развития в желательных направлениях.

Интеллект в раннем детстве

Оценка интеллекта в двух крайних точках возрастного диапазона ставит перед исследователями ряд теоретических и прикладных (связанных с интерпретацией тестовых показателей) проблем. Одна из этих проблем связана с ответом на вопрос, какие функции следует тестировать. Из чего складывается интеллект младенца и до-

школьника? Что представляет собой интеллект пожилого человека? Вторая проблема отчасти связана с первой. В отличие от школьника младенец и дошкольник не подвергаются стандартизованным последовательностям обучающих воздействий, прописанных в школьных программах. При разработке тестов для уровня начальной, средней и высшей школы у создателя теста в распоряжении имеется большой резерв эмпирического материала, на основе которого он может строить тестовые задания. Вместе с тем опыт, получаемый ребенком до поступления в школу, отличается гораздо меньшей нормативностью, несмотря на некоторое единообразие методов семейного воспитания в рамках конкретной культуры. При таких обстоятельствах конструировать тесты значительно труднее, как, впрочем, и интерпретировать их результаты. В какой-то мере трудности того же порядка возникают при тестировании пожилых людей, закончивших школу много лет назад и с тех пор занимавшихся самой разнообразной деятельностью. В данном и следующем разделах будут рассмотрены некоторые последствия этих проблем для тестирования в период раннего детства и взрослости соответственно.

Прогностическая валидность тестов для младенцев и дошкольников. Вывод, вытекающий из лонгитюдных исследований, состоит в том, что тесты для дошкольников (особенно при проведении их после достижения возраста 2 лет) имеют умеренную валидность в предсказании последующего выполнения тестов интеллекта, а тесты для младенцев фактически не обладают подобной валидностью (Bayley, 1970; Lewis, 1973; McCall, Hogarty, & Hurlburt, 1972). Объединив результаты восьми исследований, Р. МакКол и его коллеги (1972) вычислили медианы коэффициентов корреляции между показателями тестов, проведенных с детьми в первые 30 месяцев их жизни, и значениями их *IQ*, полученными в период от 3 до 18 лет. В результате обнаружилось несколько тенденций. Во-первых, тесты, предъявляемые в первый год жизни, обладают крайне низкой долгосрочной прогностической валидностью или вовсе не имеют ее. Во-вторых, тесты для младенцев обнаруживают некоторую валидность при предсказании *IQ* в период дошкольного детства (3–4 года), но корреляции резко падают к концу этого периода, когда дети достигают школьного возраста. В-третьих, после 18 мес. коэффициенты прогностической валидности достигают среднего уровня и стабилизируются, оставаясь большей частью в пределах 0,40–0,50. Когда прогнозы делаются на основании испытаний в этом возрасте, корреляции, по-видимому, будут получены примерно одного порядка, независимо от того, в какой точке возрастного диапазона от 3 до 18 лет проводится повторное тестирование.

Недостаток долгосрочной прогностической валидности у тестов для младенцев требует дополнительного анализа и оценки в свете других относящихся к этой проблеме данных. Во-первых, предсказания можно улучшить за счет учета тенденций возрастного развития, выявляемых путем неоднократного тестирования. Во-вторых, ряд исследователей обнаружили, что тесты для младенцев обладают значительно более высокой прогностической валидностью в отношении клинических популяций по сравнению с нормальными популяциями. Сообщалось о значимых коэффициентах валидности порядка 0,60–0,70 и даже выше при обследовании детей с исходным *IQ* ниже 80, а также групп детей с установленными или предполагаемыми неврологическими нарушениями (Ireton, Thwing, & Gravem, 1970; Knobloch, & Pasamanick, 1963; Werner, Honzik, & Smith, 1968). По-видимому, тесты для младенцев могут принести наибольшую пользу

в качестве вспомогательных средств при диагностике дефективного развития, вызванного органической патологией наследственного либо приобретенного характера.

При отсутствии органической патологии развитие ребенка после рождения во многом определяется той средой, в которой он растет и воспитывается. И не стоит рассчитывать на то, что тест может предсказать особенности средовых воздействий. В сущности, уровень образования родителей и другие более конкретные характеристики домашней обстановки являются более хорошими предикторами последующего интеллектуального уровня, чем показатели теста для младенцев. И все же, после 18 мес. прогностические возможности тестов для младенцев заметно улучшаются, если наряду с тестовыми показателями учитываются показатели семейного социо-экономического статуса (Bayley, 1955; McCall et al., 1972; Pinneau, 1961; Werner et al., 1968). Высказывалось также предположение, что индивидуальные различия в младенчестве могут быть относительно несущественными и преходящими, так как нормальное развитие на этой ранней стадии носит, по существу, общий для вида *Homo Sapiens* характер (R. B. McCall, 1981). В последующие годы индивидуальные развития, расширяясь, становятся все более устойчивыми в возрастном плане и дают более высокие корреляции с генетическими и средовыми факторами (Plomin, De Fries, & Fulker, 1988). Тем не менее следует отметить, что в 1990-х гг. возросло число работ, посвященных выяснению прогностической ценности когнитивного поведения младенцев, и результаты этих исследований выглядят многообещающими (Colombo, 1993).

Природа интеллекта в раннем детстве. Валидность (как обоснованность) тестов интеллекта младенцев и смысл критериев оценки выполнения таких тестов становятся более ясными и понятными на фоне исследований природы детского интеллекта. Результаты таких исследований не подтверждают концепцию устойчивой и единой интеллектуальной способности в младенчестве (Lewis, 1973, 1976; McCall et al., 1972). Пренебрежимо малые корреляции получаются даже если тестирование повторяется всего через три месяца, а корреляции с показателями по тем же или другим шкалам, полученными в 2-летнем или более старшем возрасте обычно незначимы. Кроме того, практически нет корреляций между показателями различных шкал, полученными при обследовании одних и тех же детей младенческого возраста. Эти результаты были получены как при использовании стандартизованных инструментов, таких как Шкалы развития младенцев Бейли, так и при работе с порядковыми шкалами Пиаже (Gottfried, & Brody, 1975; King, & Seegmiller, 1973; Lewis, 1976; Lewis, & McGurk, 1972).

Некоторые исследователи пришли к выводу, что несмотря на явно недостаточную прогностическую валидность тесты интеллекта младенцев являются валидными индикаторами когнитивных способностей ребенка на момент тестирования (Bayley, 1970; Stott, & Ball, 1965; Thomas, 1970). Согласно этой точке зрения, основная причина пренебрежимо малых корреляций между младенческими тестами и последующим выполнением интеллектуальных тестов — в изменении с возрастом типа и состава интеллекта. Интеллект младенца качественно отличается от интеллекта школьника и состоит из иного сочетания способностей.

Серия тщательных исследований МакКола и его сотрудников была посвящена изучению меняющейся природы младенческого интеллекта в течение первых двух лет жизни, разбитых на шестимесячные периоды (R. B. McCall, 1976; McCall, Eichorn, & Hogarty, 1977; McCall et al., 1972). Путем статистического анализа интеркорреляций между различными умениями в рамках каждого шестимесячного периода, а также

корреляций между одними и теми же умениями, проявляемыми в разные периоды, и между различными умениями, эти исследователи пытались отыскать в поведении младенцев предвестников более поздних, связанных с развитием, событий. Один вывод, напрашивающийся из данного исследования, состоит в том, что преобладающее в разные возрастные периоды поведение обнаруживает качественные изменения, которые представляют собой организованную и поддающуюся объяснению последовательность переходов. Когда реакции младенцев на задания из «Таблиц развития» Гезелла были подвергнуты факторному анализу отдельно по всем шестимесячным периодам, значения первого фактора на каждом возрастном уровне значимо коррелировали между уровнями. И это несмотря на то, что поведенческая структура этих первых факторов варьировала от возраста к возрасту. Иначе говоря, проявления умственной компетентности оказались специфичными для каждого из изученных шестимесячных периодов, хотя компетентность в одном из них предсказывала компетентность в более поздние периоды, но лишь в том случае, если каждая оценивалась по соответствующему данному возрастному периоду поведению.

Для описания изменений в соответствующих возрасту проявлениях интеллектуальной компетентности было введено понятие *возрастных трансформаций (developmental transformations)*. Дополнительные доказательства таких качественных изменений в компетентности поведения получены в исследованиях Ярроу и его коллег, посвященному изучению того, как младенец овладевает ближайшей средой (Messer et al., 1986; Morgan, & Harmon, 1984; Yarrow et al., 1983; Yarrow et al., 1984; Yarrow, & Messer, 1983). Результаты исследований показали наличие временной прогрессивной последовательности как в ряду задач, вызывающих исследовательское поведение младенца, так и в ряду вызываемых этими задачами специфических форм поведения, таких как рассматривание, манипулирование и упорство в решении задач. Например, младенец впервые обнаруживает, что может воздействовать на окружающую среду: роняет кубик, чтобы посмотреть на его падение и услышать звук удара о пол, или размахивает колокольчиком, чтобы заставить его звенеть. Позднее, процесс овладения средой проявляется в более сложных и направленных на достижение заранее поставленной цели действиях, таких как использование обходных путей или выбор подходящих средств для достижения цели — игрушки, лакомства и т. д. Благодаря идентификации таких специфических для определенного возраста форм поведения можно значительно повысить эффективность исследований как конструктной, так и прогностической валидности оценок интеллекта в раннем детстве. Важно также учитывать роль содержания знаний в использовании интеллектуальных процессов и когнитивных стратегий (Reese, 1987), — факт, получающий все большее признание в исследованиях по когнитивной психологии в целом.

Следствия для программ вмешательства. Доказательная эффективность разнообразных программ вмешательства эпохи Head Start зависит от качества конкретной программы (R. C. Collins, 1993; Haskins, 1989; Zigler, & Muenchow, 1992; Zigler, & Styfco, 1993). Предназначенные главным образом для повышения готовности к обучению у детей, живущих в неблагоприятных условиях, эти программы сильно различались как по используемым методам, так и по конечным результатам. В основном это были неудачные проекты, обреченные на провал из-за ошибок в планировании реализации или оценивания результатов. Лишь в немногих из них удалось зримо показать существенные улучшения в деятельности детей, однако эти улучшения часто носили

ограниченный и кратковременный характер. В противоположность большинству программ, нацеленных исключительно на «повышение *IQ*» с помощью расплывчато определенных методов, отдельные высококачественные проекты включали перечень четко определенных, конкретных интеллектуальных навыков, которые предполагалось улучшить, и ясное описание выбранных для достижения этой цели методик обучения. В таких случаях тщательно проведенные контрольные замеры действительно показывают существенное и устойчивое улучшение соответствующих умений и навыков. Некоторое внимание уделялось также более широкому контексту реализуемой программы, который мог включать необходимую медицинскую и социальную помощь семье. Привлечение родителей к участию в программах вмешательства оказалось особенно полезным в том, что касается оказания дополнительной помощи детям-дошкольникам в домашних условиях и обеспечения дальнейшей поддержки своим детям после окончания формального проекта (Jaynes, & Wlodkowski, 1990).

Пристального внимания требуют и программы слежения, цель которых — оценить характер и продолжительность вмешательства. Для оценивания эффективности таких проектов приходится идти на изрядные методологические ухищрения (Collins & Horn, 1991; Willett, & Sayer, 1994). Помимо обычных сомнений и споров вокруг планов проверочных экспериментов, даже идеально спланированные эксперименты могут давать ложные результаты, — как положительные, так и отрицательные, — вследствие статистических артефактов, связанных с психометрическими свойствами применяемых измерительных инструментов (Bejar, 1980). Различия в трудности либо различительной способности тестовых заданий в экспериментальной и контрольной группах или между результатами претеста и посттеста в тех же группах могут привести к некорректным выводам в отношении успешности или провала программы вмешательства. Использование тестов, в которых задания и процедуры оценивания ответов разрабатывались на основе теории «задание — ответ» (глава 7), как и тестов, приспосабливаемых к уровню возможностей каждого испытуемого (глава 10), избавляет от некоторых из этих трудностей. В последние годы возобновился интерес к созданию тщательно спланированных и строго выполняемых программ, построенных по образцу некоторых более эффективных прежних программ (R. C. Collins, 1993; Consortium, 1983; Haskins, 1989; Whimbey, 1990; Zigler, & Styfco, 1993). Кроме того, разработчики современных программ могут извлечь немалую пользу из быстро растущей базы данных, полученных в ходе исследований детского интеллекта (см., например, Horowitz, & O'Brien, 1989). Один весьма многообещающий проект посвящен воздействию родительского поведения в период первых двух лет жизни ребенка на последующую интеллектуальную деятельность детей (Hart, & Risley, 1995). Первые данные, полученные в рамках этого проекта, уже достаточно убедительно свидетельствуют о тесной связи между характером и степенью контактов родителей с детьми и интеллектуальным развитием последних.

Проблемы тестирования интеллекта взрослых

Снижение интеллекта с возрастом. Отличительной особенностью шкал Векслера для измерения интеллекта взрослых (глава 8) было использование понижающихся с возрастом норм для подсчета стандартных *IQ*. Первичные («сырые») показатели субтестов *WAIS* (и *WAIS-R*) сначала преобразуются в стандартные показатели со средним,

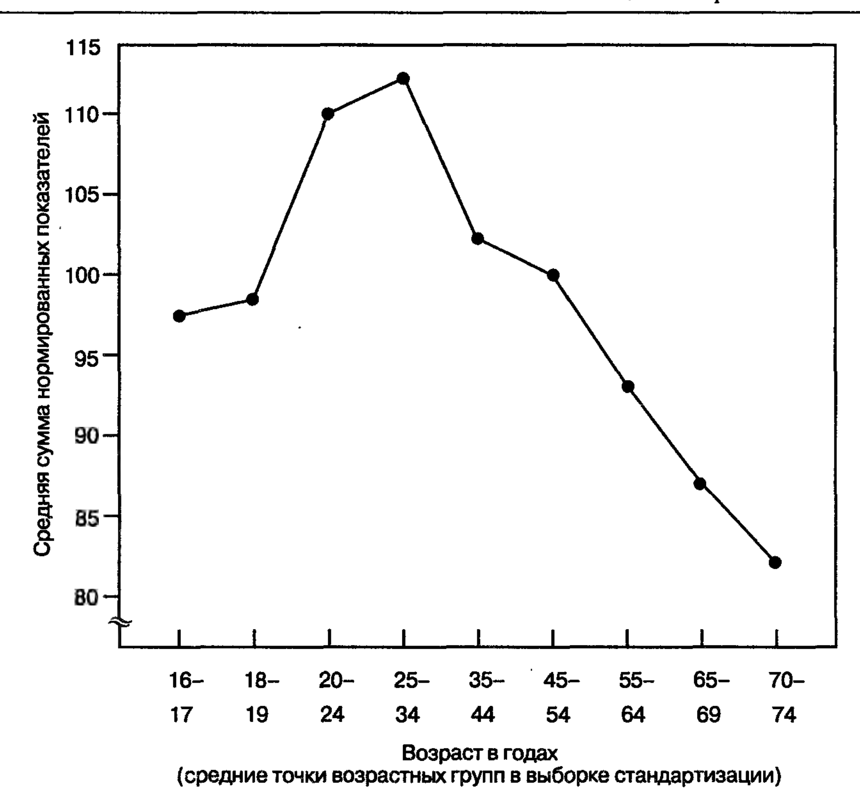


Рис. 12–1. Снижение показателей Полной шкалы *WAIS-R* с переходом от младших к старшим возрастным когортам

(По данным из Wechsler, 1981, p. 26)

равным 10, и *SD* = 3. Эти нормированные показатели выражаются в единицах фиксированных эталонных групп, составленных из 500 человек в возрасте от 20 до 34 лет, входящих в выборку стандартизации. Сумма нормированных показателей по 11 субтестам используется при определении стандартного *IQ* по таблице для соответствующего возраста. Впрочем, если просто брать суммы нормированных оценок, то можно сравнивать результаты разных возрастных групп в единицах единой непрерывной шкалы. На рис. 12–1 показаны средние этих суммарных нормированных показателей для возрастных уровней, включенных в национальную выборку стандартизации.

Как видно из рис. 12–1, показатели достигают пика между 20 и 34 годами, а затем монотонно убывают во всех более старших возрастных группах. Стандартный *IQ* находят соотнесением суммарного нормированного показателя индивидуума с нормой для его возрастной группы. Таким образом, если тестируемый обнаруживает такое же снижение уровня выполнения теста с возрастом, какое имеет место в соответствующей нормативной выборке, его *IQ* будет оставаться постоянным. Такой подход строится на исходном допущении Векслера, что вполне «нормально» для тестируемой способности любого человека снижаться с возрастом после 30 лет.

Однако для интерпретации возрастного снижения интеллекта, проиллюстрированного на рис. 12–1, мы должны учесть существенную особенность выборок, на которых проводилась стандартизация этого теста. Поскольку любая выборка стандартиза-

ции является *нормативной* выборкой, она должна отражать характеристики существующей популяции (или, используя статистический термин, генеральной совокупности) на каждом возрастном уровне (Anastasi, 1956). Из этого следует, что в случае роста образовательного уровня населения в целом на протяжении нескольких десятилетий старшие возрастные группы в любой момент времени будут находиться по сравнению с младшими на более низком уровне образования. Эти различия по уровню образования ясно отражены в выборке стандартизации *WAIS* (протестированной в 1953–1954 гг.) и *WAIS-R* (протестированной в период с 1976 до 1980 г.). В обеих выборках максимум лет обучения в учебных заведениях приходится на возрастной уровень от 20 до 34 лет, а образовательный уровень непрерывно снижается для более старших возрастных групп. Хотя выборка стандартизации *WAIS-R*, тестирование которой проводилось позднее, имела более высокий уровень образования как группа по сравнению с выборкой стандартизации *WAIS*, снижение количества лет формального образования у членов этой выборки с возрастом было столь же выраженным, как и в выборке стандартизации *WAIS*. А снижение нормированных показателей *WAIS-R* в каждой последующей возрастной когорте обнаруживает хорошее соответствие с таковым для *WAIS*.

Эти возрастные различия в объеме образования неизбежны, если мы хотим, чтобы выборка стандартизации теста была действительно репрезентативной в отношении населения страны на момент установления норм. В то же время различия в образовании затрудняют интерпретацию наблюдаемого снижения показателей: старшие возрастные группы из выборки стандартизации могут хуже выполнить данный тест не потому, что они старше, а потому, что они менее образованы по сравнению с молодыми группами.

Результаты, полученные на выборке стандартизации шкал Векслера, типичны для всех исследований интеллекта взрослых, выполненных традиционным методом поперечных срезов. Возможно, что сравнения по поперечным срезам, когда люди разных возрастов исследуются в одно и то же время, показывают кажущееся возрастное снижение, поскольку изменения в культурном уровне смешиваются с эффектами старения. Объем формального образования — это только одна из многих переменных, по которым могут различаться возрастные группы. Ведь в течение последней половины века в нашем обществе произошли и другие культурные изменения, которые делают жизненный опыт 20-летних и 70-летних совершенно непохожим. Несомненно, что изменения в области средств коммуникации (радио, телевидение, Интернет) и в области транспортных средств намного увеличили объем информации, доступный развивающемуся индивидууму. Улучшение питания и медицинской помощи также могло косвенно повлиять на развитие поведения.

Лонгитюдные исследования, основанные на повторном тестировании одних и тех же лиц в период от 5 до 40 лет, в большинстве случаев выявили противоположную тенденцию: с возрастом показатели увеличивались. Часть этих исследований проводилась с группами, отличавшимися высоким уровнем интеллекта, например с выпускниками колледжей или людьми, отобранными по критерию высокого *IQ* (Bayley, & Oden, 1955; R. B. Burns, 1966; D. P. Campbell, 1965; Nisbet, 1957; Owens, 1953, 1966). По этой причине некоторые авторы утверждали, что подобные результаты можно отнести только к людям с высоким интеллектуальным или образовательным уровнем и они неприменимы ко всей популяции. Однако сходные данные были получены в лонгитюдных исследованиях выборок нормальных взрослых со средним образованием и сред-

ним уровнем интеллекта (Charles, & James, 1964; Eisdorfer, 1963; Tuddenham, Blumenlrantz, & Wilkin, 1968), а также умственно отсталых взрослых, не помещенных в специальные интернаты (Baller, Charles, & Miller, 1967; Bell, & Zubek, 1960; Charles, 1953).

Ни исследования методом поперечных срезов, ни лонгитюдные исследования сами по себе не позволяют дать неоспоримую интерпретацию наблюдаемых возрастных изменений. С одной стороны, возрастные различия в уровне образования могут привести к мнимому возрастному снижению уровня выполнения тестов интеллекта в исследованиях, основанных на стратегии поперечных срезов. С другой стороны, по мере того как люди становятся старше, они сами подвергаются воздействию культурных изменений, могущих улучшить выполнение ими интеллектуальных тестов. Блестящий анализ методологических трудностей каждого из этих подходов, вместе с необходимыми для их реализации планами экспериментов, можно найти в целом ряде публикаций.[1] По существу, для снятия неопределенности необходимо сочетание нескольких подходов, как это имеет место в *когортно-последовательном плане (cross-sequential design)* исследования (K. W. Schaie, 1065, 1994; Shock et al., 1984). Этот план объединяет данные, получаемые в традиционном поперечном и продольном тестировании, с так называемыми *сравнениями с временной задержкой (time-lag comparisons)* или, иначе говоря, с отсроченными сравнениями однотипных результатов. Такие сравнения требуют тестирования одинаковых возрастных когорт[2] в разные периоды времени, например группа 20-летних, протестированная в 1940 г., сравнивается с группой 20-летних, протестированной в 1970 г.

Лишь немногие исследования снабжают нас данными, позволяющими провести по крайней мере частичный анализ факторов, влияющих на возрастное изменение выполнения теста. Оуэнс (Owens, 1966), повторяя тестирование бывших студентов университета штата Айова через 40 лет, а Д. Кэмпбелл (D. P. Campbell, 1965) — бывших студентов Миннесотского университета через 25 лет, протестировали еще и учившихся *в это время* на первом курсе студентов соответствующих университетов. В результате появилась возможность провести множественные сравнения результатов двух групп, протестированных в одном и том же возрасте с интервалом в 25 и 40 лет, и результатов одной группы, протестированной через те же временные интервалы. В обоих исследованиях группа бывших студентов при повторном тестировании улучшила свои результаты по сравнению с более ранним выполнением теста, выполнив тест примерно наравне с более молодой группой, впервые тестированной в более позднее время. Такие результаты свидетельствуют о том, что именно культурные изменения и другие связанные с жизненным опытом факторы, скорее чем возраст сам по себе, вызывают повышение и снижение показателей, полученных при использовании более ограниченных экспериментальных планов. Растущий интерес к исследованию научения у пожилых людей отчетливо проступает в сравнительно недавно опубликованном исчерпывающем обзоре (Kausler, 1994), определенно помогающем рассеять стереотипы о влиянии старения на научение. Хотя указанная работа основана на твердо установленных научных данных, в ней показаны многочисленные связи этих данных с событиями повседневной жизни.

---

[1] См., например, Baltes (1968), Botwinik (1984, chaps. 20, 21), Buss (1973), Nesselroade, & Reese (1973), Nesselroade, & Von Eye (1985), K. W. Schaie (1973, 1988a), Schaie & Hertzog (1986).

[2] В этом контексте под когортой понимается группа лиц одного возраста (родившихся в одном году или в какой-то определенный период времени).

Сиэтлское лонгитюдное исследование (СЛИ). Примером особенно тщательно разработанной долгосрочной исследовательской программы с использованием когортно-последовательного плана является Сиэтлское лонгитюдное исследование (*Seattle Longitudinal Study* или, сокращенно, *SLS* — K. W. Schaie, 1994; Schaie & Hertzog, 1986). Приступив к реализации этой программы в 1956 г., исследователи с помощью батареи тестов способностей[1] обследовали стратифицированную случайную выборку из 500 человек, извлеченную из совокупности примерно 18 000 обладателей медицинских страховок (*members of prepaid medical plan*). Эта совокупность, если судить по данным переписи населения США, достаточно полно отражает демографическую структуру крупных городов вместе с пригородами. Выборка включала по 25 мужчин и 25 женщин на каждом (образованном с интервалом в 5 лет) возрастном уровне от 21 года до 70 лет. На последующих стадиях СЛИ было проведено шесть циклов тестирования (с 1956 по 1991 гг.). В каждом цикле тесты предъявлялись как доступным для исследователей членам исходной выборки, так и новым выборкам испытуемых, всякий раз включаемым в исследование при проведении очередного тестирования.

Базовый экспериментальный план СЛИ включал: 1) циклические лонгитюдные ретесты одних и тех же лиц; 2) поперечные сравнения разных возрастных когорт, протестированных в одно время (например, 30-летних с 50-летними, тестирование которых проводилось в 1977 г.) и 3) сравнения отдельных возрастных когорт, протестированных в разное время (например, 30-летних с 30-летними, которым тесты предъявлялись в 1963 и 1984 гг.). С помощью соответствующего статистического анализа данных, полученных в ходе такого рода сравнений, удалось выявить изменения в уровне выполнения тестов, связанные с возрастом, культурными переменами в обществе, а также с деятельностью и личным опытом конкретного человека. Как видно на рис. 12–2, результаты этого исследования показали, что возрастное снижение большинства функций начинается позже и, в общем, происходит менее резко, чем это предполагалось на основе традиционных сравнений методом поперечных срезов.

Кроме комплексного и систематического изучения давно дискутируемой проблемы влияния возраста на выполнение тестов интеллекта, СЛИ ответило еще на несколько родственных вопросов. Например, было обнаружено, что возрастные изменения различаются в зависимости от оцениваемой функции, например, вербальной способности, числовой способности и перцептивной скорости (см. рис. 12–2). Поэтому общая мера интеллекта (такая, как *IQ*) обычно затемняла и лишала четкости действительные повышения и понижения способности. Значительная часть исследования была посвящена выяснению причин таких изменений, в особенности причин снижения результатов у пожилых людей. Среди главных причин снижения уровня выполнения интеллектуальных тестов оказались плохое состояние здоровья, специфические болезни, бездеятельность, отсутствие непрерывного упражнения конкретных функций и такие состояния личности, как ослабленная мотивация и пониженная гибкость. На основе полученных данных разрабатывались программы вмешательства, преследу-

---

[1] Первоначально использовалась батарея Первичных умственных способностей, разработанная Тёрстоуном на основе факторного анализа общепризнанных аспектов интеллекта (см. главу 11). Позднее, специально для данного проекта была разработана новая версия этой батареи — Тест умственных способностей взрослых Шайи—Тёрстоуна (*Schaie-Thurstone Adult Mental Abilities Test* — K. W. Schaie, 1988b). В конечном итоге для описания полученных данных были выбраны показатели, соответствующие конструктам, а не отдельным тестам, чтобы обеспечить большую устойчивость и обобщаемость результатов.

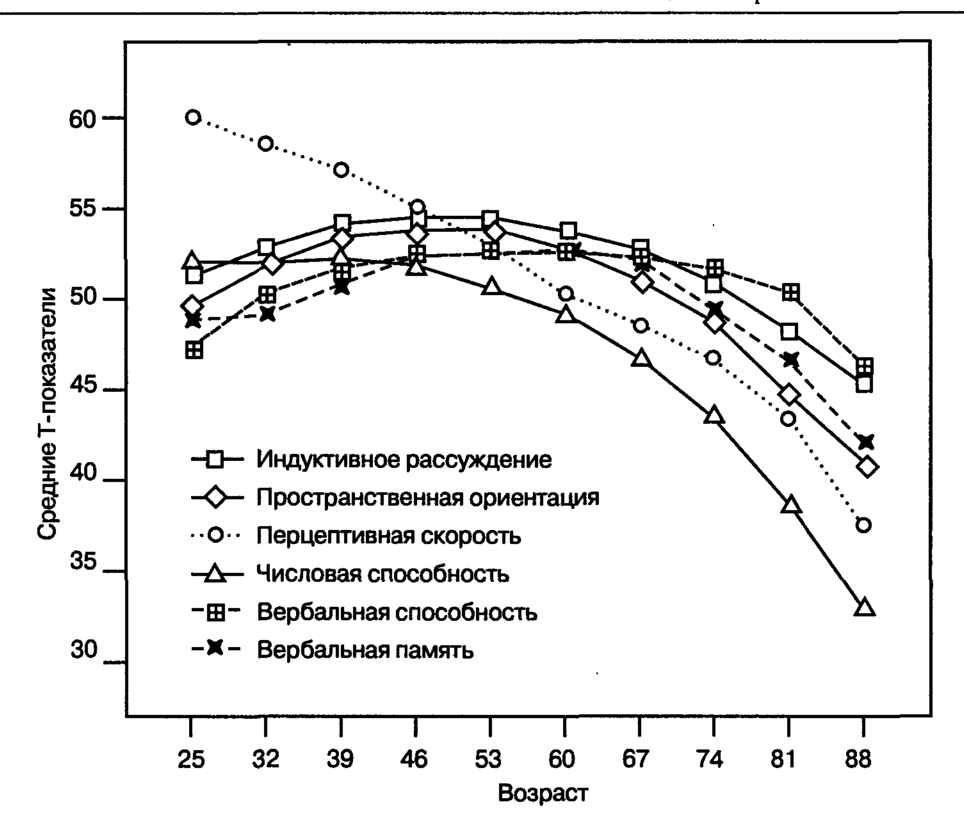


Рис. 12–2. Лонгитюдные оценки средних факторных показателей для конструктов способностей. *(Из K. W. Schaie, 1994, p. 308)*

Примечание. Данные получены в результате семилетних интраиндивидных ретестов.

(Copyright © 1994 by American Psychological Association. Воспроизводится с разрешения)

ющие цель приостановить наблюдаемое снижение конкретной способности с возрастом или даже изменить направление процесса на противоположное. Современный этап продолжающегося СЛИ ориентирован главным образом на разработку таких коррекционных программ (см. K. W. Schaie, 1994; Schaie, & Hertzog, 1986).

Индивидуальные различия и возраст. В добавление к основным результатам, согласно которым снижение интеллектуальных функций с возрастом менее выражено и наступает в жизни человека позднее, чем прежде предполагалось, современные исследования, как правило, выявляют широкие индивидуальные различия в способности на всех возрастных уровнях. Любое обобщение, относится ли оно к возрастному спаду или когортным различиям, должно смягчаться признанием широких индивидуальных различий, обнаруживаемых во всех ситуациях. Индивидуальные различия внутри любого возрастного уровня намного превосходят средние различия между двумя возрастными уровнями, вследствие чего распределения показателей у представителей разных возрастов существенно перекрываются. А это означает, что можно найти немало пожилых людей, выполняющих тесты наравне с молодыми. Более того, лучшие результаты в старших группах превосходят худшие результаты в младших груп-

пах. Такое перекрытие результатов не ограничивается только смежными возрастными уровнями, диапазоны выполнения теста перекрываются даже в тех случаях, когда сравниваются крайние возрастные группы. Так, некоторые 80-летние могут выполнить тест лучше некоторых 20-летних.

Однако для обсуждаемой темы значительно важнее то, что сами *изменения*, которые происходят с возрастом, варьируют от человека к человеку. Так, между 50 и 60 годами одни могут снизить уровень выполнения теста, другие остаться на прежнем уровне, а третьи повысить свои результаты. Степень изменения, независимо от того, будет ли это повышение или понижение, также широко варьирует у разных людей. Кроме того, углубленные исследования лиц преклонного возраста, перешедших 70, 80 и 90-летний рубеж, показывают, что функционирование интеллекта гораздо сильнее связано с состоянием здоровья конкретного человека, чем с его хронологическим возрастом (Birren et al., 1983; Palmore, 1970; Schaie, & Gribbin, 1975). К другим факторам, способствующим сохранению интеллектуального статуса, относятся благоприятная среда, с множеством возможностей стимуляции интеллектуальной деятельности, и поддержание гибкого образа жизни (K. W. Schaie, 1994; Schaie, & Hertzog, 1986).

Природа интеллекта взрослых. Если говорить о традиционном тестировании интеллекта, то оно было прежде всего ориентировано на школьников и студентов — представителей по существу одного, связанного с формальным обучением, этапа жизненного пути человека. Создатели тестов для этих уровней могут заимствовать необходимый им материал из обширного практического опыта, систематизированного в учебных программах. Большей частью интеллектуальные тесты измеряют, насколько хорошо индивидуум овладел интеллектуальными навыками, которым обучают в школе, а они, в свою очередь, могут предсказать, насколько хорошо он подготовлен к получению образования на следующем уровне. Тесты для взрослых, в том числе шкалы Векслера и примененные в Сиэтлском лонгитюдном исследовании тесты ПУС (первичных умственных способностей), строятся в основном на этом общем для многих и, главное, поддающемся учету опыте школьного образования. По мере того как человек становится старше, опыт формального обучения становится для него все более далеким прошлым, а общее ядро такого опыта, используемое разработчиками тестов интеллекта, становится все менее адекватной основой для оценки интеллектуальных функций взрослых. Занятия взрослых различаются гораздо больше, чем занятия школьников. Совокупный опыт взрослости может поэтому стимулировать у разных людей дифференциальное развитие способностей.

Поскольку тесты интеллекта тесно связаны с академическими способностями, неудивительно, что даже самые первые исследования взрослых выявили больший возрастной прирост показателей у тех, кто дольше продолжал свое образование (D. P. Campbell, 1965; Härnqvist, 1968; Husén, 1951; Lorge, 1945; Owens, 1953). Точно так же люди, род занятий которых достаточно «академичен» и связан с использованием вербальных и числовых способностей, по-видимому, на протяжении многих лет сохраняют или даже повышают свои показатели по тестам интеллекта, в то время как у людей, в чьих профессиях преобладают механические действия или межличностные контакты, эти показатели могут с годами снижаться. Предварительные данные в поддержку этой гипотезы опубликовал М. Уильямс (M. Williams, 1960), сравнивший выполнение 100 испытуемыми в возрасте от 65 до 90 с лишним лет серии вербальных и невербальных тестов. В результате было обнаружено поразительное соответствие между професси-

ей человека и его относительным выполнением этих двух типов задач. При лонгитюдном изучении взрослых также были установлены не оставляющие сомнений связи между изменениями общего *IQ* и отдельными пунктами биографических вопросников (Charles, & James, 1964; Owens, 1966).

Время и среда, предъявляя к деятельности живущего в них человека свои особые требования, благоприятствуют развитию вполне определенных навыков. На протяжении жизни эти требования меняются, и их характер различен для дошкольников, школьников, взрослых разных специальностей и удалившихся от дел стариков (Baltes, Reese, & Lipsitt, 1980). Демминг и Пресси (Demming, & Pressey, 1957) одними из первых продемострировали следствия этого факта для тестирования интеллекта. Эти исследователи начали с анализа типичных профессиональных обязанностей взрослых, провели содержательный обзор читаемого материала и собрали сведения о ежедневно осуществляемой деятельности и типах решаемых проблем. На этой основе они подготовили предварительные формы 20 тестов, «родных» для представителей старшего возраста. В тестах основной акцент делался на объеме практической информации, особенностях суждений и социальной перцепции. Результаты трех из этих тестов, проведенных вместе со стандартными вербальными и невербальными тестами на выборках разного возраста, показали, что в новых тестах лица более старшего возраста превосходили более молодых, тогда как для традиционных тестов сохранялось обратное соотношение. Все эти типы исследований свидетельствуют об одном: будет ли в период взрослости иметь место повышение или снижение тестовых показателей по мере увеличения возраста людей, во многом зависит от того, какой опыт приобрел человек за эти годы, и от степени связи между этим опытом и функциями, охватываемыми данными тестами. На более широком уровне обобщения можно сказать, что все результаты тестирования могут быть наилучшим образом проинтерпретированы в рамках соответствующего контекста. Функционально-содержательный анализ поведения в разнообразных сферах жизни взрослых людей, включая профессиональную и другие сферы социально значимой деятельности, должен привести как к усовершенствованию конструкции тестов, так и к прояснению смысла тестовых показателей (Anastasi, 1986b). Примечательно, что как 1980-е, так и 1990-е гг. отмечены явным ростом исследований, посвященных проблемам взрослости, особенно последнего отрезка времени жизни.[1] Психология развития человека на протяжении всей его жизни (*Life-span developmental psychology*) — быстро развивающаяся область исследований (Rutter, & Rutter, 1993).

Изменение показателей тестов интеллекта на уровне популяции

Повышение показателей. Что происходит с результатами тестов интеллекта на уровне популяции в течение продолжительных периодов времени? С этим вопросом мы уже сталкивались в связи с рассмотрением нескольких проблем. В предыдущем

[1] См., например, Bengtson, & Schaie (1989), Birren, & Bengtson (1988), Birren, & Schaie (1991), Craik, & Salthouse (1992), Fiske, & Chiriboga (1990), Haplin, & Panek (1993), Kausler (1994), Nadien (1989), Sonderegger (1992), Schulz, & Ewen (1993), Willis, & Schaie (1986).

разделе было показано, что вместе с ростом образовательного уровня населения на протяжении нескольких десятилетий соответственно повышался и средний уровень выполнения интеллектуальных тестов. В результате, более старшие, но и менее образованные, в среднем, члены нормативной выборки, получали по тесту интеллекта показатели ниже, чем более молодые, но, в среднем, лучше образованные члены этой выборки. Аналогичный феномен обнаруживается при изучении выборки стандартизации тестов для детей. При рестандартизации шкал Стэнфорд–Бине и *WISC* результаты выполнения тестов в более поздних выборках стандартизации существенно лучше, чем в более ранних выборках. Как следствие, любой конкретный ребенок получил бы более низкий *IQ* при его тестировании с помощью пересмотренной шкалы, чем в случае применения старой шкалы, причем просто потому, что его результат оценивался бы относительно более высоких норм. К тому же более высокий уровень образования родителей тех детей, которые вошли в более позднюю выборку стандартизации, был одним из условий, упомянутых в связи с повышением оцениваемого тестами уровня интеллекта.

Этот тип сравнения можно определить как лонгитюдное изучение популяций. Обычное применение лонгитюдного метода в психологических исследованиях связано с повторным тестированием одних и тех же лиц на протяжении какого-то периода времени. Однако при лонгитюдном изучении популяций определенная популяция выборочно обследуется в различные периоды времени. В этом случае сравнение производится между когортами лиц, родившихся в разное время, но протестированных в одном возрасте.[1] Несколько крупномасштабных исследований, поведенных в течение первых пяти десятилетий XX столетия, показали повышение интеллекта популяции в том виде, как он измеряется стандартизованными тестами интеллекта (Anastasi, 1985d, p. 126–130). При росте грамотности, повышении образовательного уровня и других культурных переменах не столь уж неожиданно было обнаружить, что средний уровень тестируемого интеллекта всего населения устойчиво повышался на протяжении нескольких десятилетий

В этих сравнительных исследованиях использовались разнообразные методы их организации. В одном случае один и тот же тест проводился через какой-то промежуток времени, как при обследовании 11-летних шотландских детей в 1932 и 1947 гг. (Scottish Council, 1949). В другом — репрезентативной выборке лиц давалось два теста с целью установить соответствие между двумя множествами показателей и таким образом обеспечить «перевод» результатов одного теста в результаты другого. Это было сделано при сравнении результатов тестирования солдат американской армии в период Первой и Второй мировых войн, которых обследовали с помощью Армейского альфа и Армейского общего классификационного тестов (*Army General Classification Test*) соответственно (Tuddenhaim, 1948). Третий, технически более совершенный, подход основан на создании абсолютной, независимой от выборки шкалы показателей посредством использования анкерных заданий, как было сделано при разработке тестов Совета колледжей. Применение теории «задание — ответ» (глава 7) представляет собой дальнейшее усовершенствование этого подхода.

[1] Специальное приложение этого общего метода можно распознать в «сравнении с временной задержкой», включенном К. У. Шайи (K. W. Schaie, 1965) в его упоминавшийся выше когортно-последовательный план исследования.

Понижение показателей. Будут ли показатели определенной популяции по тестам интеллекта со временем повышаться, понижаться или оставаться стабильными, зависит от многих условий. Охватываемый *период времени*, с сопутствующими ему культурными переменами, безусловно является главным фактором. *Возраст* обследуемых также имеет значение. Например, повышение образовательного уровня населения будет прямо сказываться на выполнении тестов взрослыми людьми и только косвенно — на показателях детей, так как дети в сравниваемых выборках получили на момент тестирования одинаковый объем образования. Еще одно важное условие, которое нужно учитывать, особенно при изучении специально отобранных субпопуляций, — это *коэффициент отбора (degree of selection)* в разные периоды времени. К примеру, если в 1960 г. среднюю школу посещала большая доля населения, чем в 1910, как это и было в действительности, то ученики средней школы 1910 г. представляют собой выборку из генеральной совокупности, извлеченную на основе более высокого критерия, чем выборка учеников средней школы 1960 г. Видимые противоречия между повышением и понижением показателей могут быть обусловлены характером используемых тестов, спецификой тестируемых субпопуляций (например, ориентированные на поступление в колледж старшеклассники, все взрослое население, ученики начальной школы) или специфическими периодами времени, охватываемыми исследованием (например, Flynn, 1984, 1987).

Количество и сложность условий, которые могут вызывать повышение или снижение интеллектуального уровня популяции, определяемого с помощью тестов, иллюстрируется анализом широко известного снижения показателей по Тесту академических способностей (*SAT*) Совета колледжей (Donlon, 1984, p. 188–191; Wirtz, 1977). В период между 1963 и 1977 гг. средний Вербальный показатель *SAT* упал с 478 до 429, а средний Математический показатель *SAT* снизился с 502 до 470. Чтобы понять причины этого неуклонно продолжавшегося 14 лет спада, специально назначенная комиссия заказала 38 исследований специалистам в разных областях и рассмотрела внушительное множество причинных гипотез.

Главный вывод, к которому пришла комиссия, заключался в том, что характер причин изучаемого явления существенно различался в первой и второй половинах 14-летнего периода. На протяжении первых 7 лет снижение показателей происходило преимущественно в результате изменения состава сдающих *SAT*. Из-за непрерывного роста доли выпускников средней школы, намеревающихся поступить в колледж в течение этого периода, данная выборка становилась все менее отсортированной по когнитивным навыкам, измеряемым данным тестом. Однако во втором 7-летнем периоде популяция поступающих в колледж практически стабилизировалась, и теперь особенностями выборки объяснялась гораздо меньшая часть снижения показателей. Для этого периода объяснение пришлось искать главным образом в условиях, связанных с семьей, школой и обществом в целом. Комиссия отметила, что имеющиеся данные не позволяют определить относительный вклад различных культурных изменений в снижение показателей теста. Тем не менее среди многих факторов, называвшихся в числе возможных значимых условий такого снижения, были и такие, как недооценка значения учебных стандартов, инфляция школьных отметок и автоматический (без экзаменов) перевод в следующий класс, сведение к минимуму домашних заданий, рост числа прогулов школьных занятий, все меньшее уделение внимания овладению навыками и знаниями, чрезмерное увлечение просмотром телепередач, а

также социальные потрясения конкретного исторического периода, препятствующие должному вниманию к жизни школьников.[1]

Последующий анализ (Turnbull, 1985) дал возможность предположить, что снижение показателей *SAT* в течение второго 7-летнего периода было, отчасти, отсроченным и косвенным следствием изменений в составе поступавших в колледжи на протяжении первых 7 лет. Поскольку большая доля плохо подготовленных учеников оставалась в средней школе (и подавала заявления о приеме в колледж), многие изменения условий школьного обучения можно рассматривать как приспособительную реакцию школ на возросшую разнородность своих учащихся. Подобные реакции, ведущие к снижению уровня требований школьной программы, иллюстрируются инфляцией отметок, быстрым увеличением числа факультативных курсов по профессиональным и даже по общеобразовательным дисциплинам, упрощением учебников и сокращением домашних заданий. Таким образом изменения в популяции учащихся привели к изменениям учебных программ, что, в свою очередь, повлекло за собой снижение тестовых показателей. Эта гипотеза согласуется с данными о том, что высоких показателей стало меньше в 1970-х гг. Показатели учеников, занимающих высокое место в своем классе, продолжали снижаться, тогда как показатели учеников, занимающих последние места в списке класса, стабилизировались или даже повысились. Более того, в конце 1970-х и в течение 1980-х гг. произошел перелом в уровне требований школ к обучению, который отразился в росте показателей *SAT*.

Общий обзор. О методологических проблемах, встречающихся на пути тех, кто пытается оценить изменения популяции, ясно свидетельствует неудавшийся обзор публикаций, посвященных улучшению результатов интеллектуальных тестов у населения 14 стран (Flynn, 1987). Опубликованные данные оказались настолько противоречивыми, а их объяснения такими расплывчатыми, что склонили автора обзора не делать никаких выводов, за исключением того, что тесты интеллекта на самом деле измеряют все что угодно, только не интеллект! Попытки измерить изменения популяции пока носят поисковый характер, и оптимальный способ проведения таких измерений еще предстоит разработать. Прекрасное рассмотрение методических вопросов, связанных с измерением изменений популяции, можно найти в материалах конференции по этой теме (Collins, & Horn, 1991).

Пока же для правильного понимания результатов исследований повышения и снижения показателей тестов в популяциях требуется дополнительная информация нескольких видов. Во-первых, должны быть точно описаны проводимые тесты, с уделением особого внимания тому, какие специфические процессы и какое содержание они охватывают (например, перцептивную скорость, память, вербальное понимание, пространственную ориентацию), и указаны источники получения их норм. Во-вторых, необходимо сообщать даты проведения первичного и всех последующих сеансов тестирования. В-третьих, должна приводиться релевантная информация о выборочно обследуемых популяциях, а также о любых изменениях отбора при повторном тестировании, таких, например, как потеря лиц с первоначально лучшими и худшими результатами в последующих выборках. В-четвертых, следует иметь доступ к информа-

[1] Хотя наиболее полно исследовались причины снижения показателей *SAT*, аналогичный спад показателей был отмечен и в других тестах для абитуриентов, таких как тесты из программы *ACT* (*American College Testing*), а также на уровнях средней и начальной школы.

ции о любых культурных изменениях, затрагивающих изучаемую популяцию; к ним можно отнести объем и характер образования, достижения в области средств связи и транспортных средств, могущие повлиять на межкультурные контакты, или любые другие события, затрагивающие течение жизни конкретных людей и могущие изменить степень или направление их интеллектуального прогресса.

Дополнительную путаницу, затрудняющую сравнительный обзор изменений в различных популяциях, может также вносить ошибочное, но, к сожалению, распространенное употребление термина «интеллект» (и особенно *IQ*), как если бы он означал единое, идентифицируемое свойство организма (см., например, Flynn, 1987). Если же вместо этого под интеллектом понимать объединение способностей, необходимое для эффективной деятельности и продвижения вперед в определенной среде (см. главу 11), тогда правильная интерпретация как индивидуальных показателей теста интеллекта, так и средних результатов популяций, обследованных в разное время и в разных местах, безусловно, требует знания важнейших условий, упоминавшихся выше. В настоящее время отмечается растущее признание технических проблем, встречающихся при измерении изменения популяции, а также многообразия методов оценки таких изменений в различных контекстах и с разными целями (см. особенно Gottman, 1995).

Культурное разнообразие

Применение тестов к представителям различных культур рассматривается под разными углами зрения в разных частях этой книги. В главе 18 затрагиваются социальные и этические аспекты такого тестирования, особенно в отношении групп меньшинств в составе более широкой национальной культуры. Технические проблемы, связанные с систематической ошибкой теста и систематической ошибкой задания, проанализированы соответственно в главах 6 и 7. А в главе 9 рассмотрены типичные тесты, первоначально предназначавшиеся для применения в самых разных культурах (так называемые «культурно-свободные тесты»). В этом разделе мы познакомим читателя с основными теоретическими вопросами о роли культуры в поведении, делая особый акцент на интерпретации и использовании показателей тестов интеллекта.

Область культурной психологии. В последние три десятилетия XX века наблюдался заметный рост исследований и публикаций по культурной психологии (Berman, 1990; Irvine, & Berry, 1988). Было даже проведено несколько международных конференций, посвященных почти целиком этой теме (см., например, Brislin, 1993; Cronbach, & Drenth, 1972; Manoleas, 1995). Эта область психологии рассматривает по существу поведенческие различия между группами, выросшими и функционирующими в объективно различимых культурных средах. Такие среды могут быть и узкими, как, например, квартал или деревня, и широкими, как страна или континент. Кроме того, некоторые широко определяемые культуры, такие как латиноамериканская, состоят из субкультур — мексиканской, кубинской, пуэрто-риканской, центрально- и южноамериканской, которые достаточно различаются для того, чтобы выделять их как таковые при необходимости понять индивидуальное поведение (см., например, Geisinger, 1992; Marin & Marin, 1991).

Роль культуры в поведении человека можно представить себе как форму предметной специфичности (*domain specificity*), аналогичную признаваемой в когнитивной

психологии (см. главу 11). Приступив к анализу основных психологических процессов, таких как научение, запоминание, решение задач (*problem solving*) и эмоции, когнитивные психологи вскоре обнаружили, что данные процессы проявляются в поведении, зависящем от предметной области. Например, память, решение задач и рассуждение могут очень существенно различаться в тех случаях, когда человек играет в шахматы, раздумывает над математической задачей или пишет реферат.

Культурная психология началась с изучения поведения в далеко отстоящих друг от друга и малоизвестных культурах, которые заметно отличались от собственной культуры исследователя, и быстро развилась в область систематического исследования непохожих историй жизни людей, воспитанных в разных культурах. По существу, область современной культурной психологии символизирует признание культурной специфичности всего человеческого поведения, в силу чего основные психологические процессы могут приводить к сильно различающимся действиям, аттитюдам, представлениям о себе и о мире у представителей разных культур (L. L. Adler, & Gielen, 1994; Berry, Poortinga, Segall, & Dasen, 1992; Diaz-Guerrero, 1990; Shweder, & Sullivan, 1993). Вклад культуры в поведение человека все больше осознается и интегрируется во все области психологии, от исследований и теории развитии человека на протяжении всей его жизни, социального поведения, эмоций или мышления,[1] с одной стороны, до практики индустриально-организационной, клинической и консультирующей психологии с другой.[2]

Растущее осознание роли культурных факторов во всех областях психологии нашло отражение в программе одного из ежегодных съездов Американской психологической ассоциации (APA, 1994). В добавление к однодневному семинару (в рамках программы непрерывного образования) по теме «Восприимчивость к культурным различиям при оценивании и вмешательстве», ежегодные лекции ведущих специалистов (*Master Lectures*) были посвящены общей теме: «Международные перспективы кросс-культурной психологии». Сами эти лекции были частью тематического минисъезда (*Topical Miniconvention*), посвященного теме «Культурное разнообразие: Будущее Америки» и продолжавшегося все пять дней работы съезда АПА. Дополнительным свидетельством быстро растущей кросс-культурной ориентации психологии служит основание нового журнала — *Culture and Psychology* (1995).

Культурные различия против культурной отсталости. Когда психологи в первой четверти XX столетия начали разрабатывать инструменты для кросс-культурного тестирования, они надеялись на то, что есть хотя бы теоретическая возможность измерить «наследственный интеллектуальный потенциал» независимо от влияния культуры. Поведение индивидуума представлялось покрытым своего рода культурной облицовкой, проникнуть под которую предполагалось с помощью того, что в то время называли «культурно-свободными» (*culture-free*) тестами. Последующее развитие генетики и психологии показало ошибочность этой концепции. Теперь мы понимаем, что факторы наследственности и среды взаимодействуют на всех стадиях развития организ-

---

[1] См., например, Gormly, & Brodzinsky (1993), Kitayama, & Marcus (1994), Mistry, & Rogoff (1985), Nugent, Lester, & Brazelton (1991), Rogoff (1990), Rogoff, & Chavajay (1995), Smith, & Bond (1993), Topping, Crowell, & Kobayashi (1989).

[2] См., например, Freilich, Raybeck, & Savishinsky (1991), Pedersen (1987), Pedersen, & Ivey (1993), Triandis, Dunnette, & Hough (1994).

ма и что их совместное влияние сложно переплетено в фактическом поведении индивида. Культурой пропитаны почти все контакты человека со средой. И поскольку всякое поведение подвержено влиянию культурной среды, в которой индивидуум воспитывается, а психологические тесты есть не больше чем выборочная проверка поведения, культурные влияния будут и должны сказываться на выполнении теста. Тщетно поэтому пытаться изобрести тест, *свободный* от влияния культуры. Позднее эта цель была изменена на другую — создать тесты, основанные исключительно на *общем* для разных культур опыте. Вот почему такие термины, как «общекультурные» (*culture-common*), «культурно-беспристрастные» (*culture-fair*) и «кросс-культурные» (*cross-cultural*) тесты, заменили собой более старый термин «культурно-свободные» тесты.

И все же ни один тест не может быть одинаково применимым или равно «справедливым» для всех культур. Множество параметров, по которым различаются культуры, столь же велико, как и множество разнообразных культурно-беспристрастных тестов. Тест, не предполагающий умения читать, может оказаться культурно-беспристрастным в одной ситуации, неязыковый тест — в другой, тест действия — в третьей, а адаптация переведенного на другой язык вербального теста — в четвертой. Многочисленные варианты имеющихся в наличии кросс-культурных тестов не являются взаимозаменяемыми, однако они полезны при проведении разных типов межкультурных сравнений. Более того, маловероятно, что какой-либо тест может быть одинаково «справедливым» более чем для одной культурной группы, особенно если культуры совершенно несхожи. Сокращая проявление культурных различий в выполнении теста, кросс-культурные тесты не могут устранить их полностью. Каждый тест ставит в более благоприятные условия представителей той культуры, в которой он создавался. Простое использование бумаги и карандаша или предъявление абстрактных задач, не имеющих непосредственного практического значения, будут благоприятствовать одним культурным группам и мешать другим. Эмоциональные и мотивационные факторы также влияют на выполнение теста. Среди множества релевантных условий, различающихся при переходе от одной культуре к другой, можно упомянуть естественный интерес к содержанию теста, раппорт с тестирующим, стремление хорошо выполнить тест, желание превзойти других и сложившиеся традиции индивидуального или совместного решении задач (*problem solving*).

Культурные различия становятся культурными барьерами, когда человек покидает ту культуру или субкультуру, в которой он воспитывался, и пытается действовать, конкурировать и добиваться успеха в другой культуре. С более широкой точки зрения, однако, именно эти контакты и взаимообмены между культурами стимулируют развитие цивилизаций. Культурная изоляция, хотя, возможно, в чем-то более комфортна для отдельных людей, ведет к застою в развитии общества.

Близкое понятие — культурная депривация. Хотя этот термин использовался в самых разных смыслах, Фейерстейн (Feuerstein, 1980, 1991; Feuerstein, & Feuerstein, 1991) наделил термин «культурная депривация» особым значением и сделал центральным в своей программе когнитивного тренинга. Он рассматривал культурную депривацию как состояние пониженной когнитивной модифицируемости, вызванное недостатком *опыта опосредствованного обучения (mediated learning experience)*. Передача накопленных в культуре знаний от одного поколения к другому — это чисто человеческое явление. В этом процессе родитель или другое лицо, выполняющее функции воспитателя, действует как *посредник (mediating agent)* в том, что касается отбора и организации стимулов, с которыми сталкивается ребенок. Фейерстейн считает

такое опосредствованное обучение существенным для когнитивного развития ребенка, поскольку оно благоприятствует формированию и закреплению учебных установок, ориентаций и других моделей поведения, облегчающих последующее обучение. У детей, которые по какой-то причине не смогли приобрести опыта такого опосредствованного обучения, отсутствуют предпосылки для высшей познавательной деятельности. В противоположность этому, те, кто получил опыт опосредствованного обучения в рамках своей культуры, развили определенные навыки и привычки, составляющие необходимое условие непрерывной модифицируемости, и могут приспособиться к требованиям новой культуры после относительно короткого переходного периода. Кроме того, вполне вероятно, что в развивающихся странах психологи из числа коренного населения со временем разработают и будут использовать тесты, соответствующие родной культуре.

Согласно другой точке зрения, *культурные стереотипы* просто в силу своего существования могут прямо влиять на выполнение теста конкретным человеком (Steele, Spencer, & Aronson, 1995). В одном долгосрочном исследовании было установлено, что знание о существующих стереотипах может влиять на мотивацию и отношение к тесту некоторых испытуемых, проявляясь их в отвлекаемости, снижении самооценки и усилий, а также в слабой надежде на успешный результат. Такая реакция получила название *незащищенности от стереотипов (stereotype vulnerability)*, и было обнаружено ее влияние на тестовые показатели как в гендерных, так и в этнических сравнениях. Для того чтобы программы работы с особо защищаемыми группами населения были эффективными, требуется больше, чем одно только специальное коррективное обучение. По крайней мере, в отдельных случаях ожидание определенного результата в конкретных областях тестирования, таких как вербальная, математическая или пространственная, может потребовать особого внимания, чтобы предотвратить неудачу тех, кто при других обстоятельствах мог бы выполнить соответствующие задания на среднем или даже самом высоком уровне.

Язык в транскультуральном тестировании. Большинство традиционных кросскультурных тестов построены на невербальном содержании в надежде получить по возможности более беспристрастную к культуре мерку для тех же интеллектуальных функций, которые измеряются вербальными тестами интеллекта (см. главу 9). Оба допущения, лежащие в основе этого подхода, вызывают сомнения. Во-первых, невозможно согласиться с тем, что невербальные тесты измеряют те же функции, что и вербальные, даже если эти тесты кажутся очень похожими. Тест пространственных аналогий — это не просто невербальный вариант теста словесных аналогий. Некоторые из первых неязыковых тестов, таких как Армейский бета, были сильно нагружены перцептивными способностями и способностью оперировать пространственными представлениями, совершенно не связанными с вербальными и числовыми способностями. Даже в тестах типа Прогрессивных матриц Равена и в других неязыковых тестах, специально предназначенных для выявления «чистой» способности к логическому выводу и абстрактной концептуализации, факторный анализ обнаружил сильный вклад невербальных факторов в дисперсию тестовых показателей (например, R. S. Das, 1963). Такие результаты подтверждаются более современными исследованиями в когнитивной психологии, неоднократно демонстрировавшими *зависмость процессов мышления от содержания (content specificity)*. Стратегии и навыки решения задач (*problem solving*) развиваются в процессе реагирования на содержание специфической предметной области и в специфических контекстах (см. главу 11).

Если посмотреть с другой точки зрения, все большее количество фактов говорит о том, что неязыковые тесты могут быть больше нагружены культурными факторами, чем языковые. Исследования разнообразных культурных групп во многих странах мира выявили большие групповые различия в выполнении целого ряда невербальных тестов по сравнению с вербальными (Irvine, 1969a, 1969b; Jensen, 1968; Ortar, 1963, 1972; Trimble, Lonner, & Boucher, 1983; Vernon, 1969). Есть и данные о том, что рисуночные (*figural*) тесты, возможно, больше подвержены эффектам обучения, чем вербальные и числовые (Irvine, 1983). Использование графической формы представления заданий само по себе может оказаться непригодным в культурах, представители которых не привыкли пользоваться символическими рисунками. Двумерное воспроизведение предмета не является его точной копией, в нем даны лишь некоторые признаки, которые, вследствие прошлого опыта, приводят к восприятию этого предмета. Если такие признаки сильно редуцированы, как в рисунке головы, служащей символическим изображением человека, то при отсутствии необходимого прошлого опыта правильное восприятие может и не последовать. Накопленные к настоящему времени эмпирические данные указывают на заметные различия в восприятии рисунков представителями разных культур (R. J. Miller, 1973; Segall, Campbell, & Herskovits, 1966).

Согласно еще одной точке зрения, невербальные, пространственно-перцептивные (*spatial-perceptual*) тесты часто требуют сравнительно абстрактных мыслительных процессов и аналитических когнитивных стилей, характерных для представителей среднего класса западных культур (J. W. Berry, 1972; R. A. Kohen, 1969). Людям, воспитанным в иной культурной среде, такие подходы к решению задач могут быть менее привычны. Культуры различаются по тому значению, какое они придают обобщению и поиску общих признаков непохожих впечатлений. В некоторых культурах поведение типично связывается с конкретными контекстами и ситуациями. Ответ на вопрос может зависеть от того, кто его задает, и от типа затрагиваемого им содержания (Cole, & Bruner, 1971; Goodnow, 1976; Neisser, 1976, 1979).

Конечно, методические трудности в проведении вербального теста в культурах, говорящих на одном языке, отсутствуют, но когда языки различны, тест нуждается в переводе, и здесь возникают проблемы сопоставимости норм и эквивалентности показателей. Следует также отметить, что простого перевода редко бывает достаточно. Обычно требуется некоторая адаптация и пересмотр содержания заданий, поскольку оно может быть более знакомым представителям одной культуры, чем другой. На выполение теста могут существенно влиять даже более тонкие различия. Например, относительная длина слова или звуковое сходство между разными словами в одном из языков может изменить трудность чтения заданий теста при переводе его с одного языка на другой (Valencia, & Rankin, 1985). Вследствие множества аспектов, в которых переводные версии теста могут отличаться от оригинала, вряд ли разумно предполагать их сравнимость (Duran, 1989; Marín & Marín, 1991). *Стандарты тестирования* (AERA, APA, NCME, 1985, chap. 13) предписывают, что надежность, валидность и нормы переводных версий теста должны независимо устанавливаться для любой популяции, в которой предполагается их использование.

Следует иметь в виду, что культурные факторы, влияющие на ответы тестируемого, вероятно, сказываются и на более широкой области поведения, для выборочной проверки которого создается данный тест. В англоязычной культуре, например, недостаточное владение английским языком может мешать ребенку не только при выполнении теста интеллекта, но также быть помехой в учебе, общении с одноклассниками

и игровой деятельности, препятствуя тем самым нормальному ходу интеллектуального и эмоционального развития. У взрослого слабое знание языка культуры, в которой он живет, могло бы серьезно ограничить уровень выполнения работы, межличностные отношения и другие виды жизненно важной активности. Можно привести множество других примеров таких культурных различий. Одни относятся к когнитивным различиям, таким как трудности чтения или неэффективные стратегии решения абстрактных задач; другие — к различиям в аттитюдах и мотивации, таким как отсутствие интереса к интеллектуальным занятиям, враждебность к представителям власти, низкая потребность в достижениях или низкая самооценка. Все подобные состояния поддаются исправлению и улучшению с помощью целого ряда средств, от обучения языку и повышения уровня грамотности до личного консультирования и психотерапии. И все они, по всей вероятности, влияют как на выполнение тестов, так и на повседневную деятельность ребенка и взрослого.

Важная роль языка в выполнении тестов и в повседневной деятельности стимулировала разработку тестов владения языком (*language proficiency tests*), причем как языка страны рождения, так и языка страны проживания. Большинство ныне доступных в США тестов такого рода проверяют владение английским и испанским языками. Перечень тестов владения языком, а также ссылки на литературу, освещающую исследования, разработку и оценку таких тестов, можно найти в работе Duran (1989, p. 574–577). Что касается более широкой трактовки билингвизма, см. de Groot & Barry (1993).

Ситуация тестирования. Быстрое расширение транскультуральных контактов в современном мире повышает вероятность предъявления тестов выходцам из различных культур. Каждый, кто проводит тестирование, может ожидать, что ему придется столкнуться хотя бы с одним тестируемым из культуры, отличающейся от его собственной. Следовательно, подготовка специалистов в этой области должна включать основные сведения об одной или большем числе разнородных культур, с уделением особого внимания вероятным культурным влияниям на развитие индивидуального поведения. Еще более важными являются возможные влияния таких различий на отношение тестируемого к тестированию. Примерами некоторых общих источников различного поведения в ситуации тестирования могут служить вариации в представлении испытуемых о себе, их мировоззрении, степени самораскрытия и привычки решать задачи в одиночку или в группе.

Как уже было показано ранее в этой главе, простое удаление из теста тех частей, которые расцениваются как особенно трудные вследствие иного культурного происхождения индивидуума, привело бы только к снижению его прогностической валидности и лишило бы возможности привлечь внимание тестируемого к тем областям, которые необходимо усилить для эффективной деятельности в ожидаемой обстановке. В соответствии с этим, традиционный подход заменяется решением проблемы путем перемещения фокуса на *поведение тестирующего* в ситуации тестирования.

Восьмидесятые и девяностые годы принесли множество публикаций (руководств, инструкций, журнальных статей и т. д.), посвященных подготовке и должному поведению специалистов, проводящих тестирование представителей других культур (Atkinson, Morten, & Sue, 1993; Myers, Wohlford, Guzman, & Echemendia, 1991; Stricker et al., 1990). Часть публикаций была посвящены конкретным проблемам тестирования студентов (Samuda, Kong, Cummins, Lewis, & Pascual-Leone, 1991), другая их часть основ-

ное внимание уделяла работе с детьми (Miller-Jones, 1989; Rogoff, & Morelli, 1989), третья касалась вопросов тестирования конкретных культурных совокупностей или их более мелких образований (Dana, 1984; Diaz-Guerrero, & Szalay, 1991), однако в подавляющем большинстве этих публикаций рассматривалось кросс-культурное тестирование как общая проблема. Их диапазон был весьма широк — от самых общих и кратко сформулированных руководящих принципов, опубликованных Американской психологической ассоциацией в виде небольшой брошюры в 1991 г. и перепечатанных в журнале *American Psychologist* в 1993 г. («Guidelines», 1993), до всестороннего и подробного рассмотрения вопросов тестирования представителей других культур в книге Роберта Даны (Dana, 1993). Два последних источника следует проштудировать каждому, кто планирует проводить тестирование. Здесь также уместно отметить, что всестороннее обсуждение разновидностей, проблем и преимуществ билингвизма можно найти в статье Де Гроота и Бэрри (de Groot, & Barry, 1993).

По существу, особая роль тестирующего в кросс-культурном тестировании включает, во-первых, получение полных сведений (в предварительной встрече с тестируемым), касающихся культурной идентичности, степени и типа аккультурации, а также особенностей исходной культуры, могущих, предположительно, повлиять на выполнение теста данным человеком. Во-вторых, тестирующему необходимо приспособить свое поведение к потребностям конкретного тестируемого. В этой связи тестирующий должен обдумать ряд вопросов: с чего начать тестирование? как объяснить цель теста? как мотивировать тестируемого выполнить тест надлежащим образом? Кроме того, поведение самого тестирующего должно способствовать установлению доброжелательных отношений с тестируемым. При интерпретации и дальнейшем использовании результатов тестирования безусловно следует учитывать культурные факторы, равно как и характер обратной связи, т. е. в какой форме и кому предполагается сообщить эти результаты.

Ричард Дана (Dana, 1993) разработал набор вопросников, помогающих тестирующему в получении необходимой информации от тестируемых. Некоторые из этих вопросников предназначены для общего употребления. Другие же были составлены для специфических культур (североамериканских индейцев, американцев азиатского происхождения) или для отдельных субкультур внутри них. В качестве заключительного замечания следует добавить, что акцент на важности получения хоть *каких-то* релевантных сведений о жизненной истории тестируемого для правильного понимания и использования тестовых показателей привлекает внимание к желательности подобной практики в тестировании *любого* индивидуума, независимо от культурных различий.

Часть 4

ТЕСТИРОВАНИЕ ЛИЧНОСТИ



13 СТАНДАРТИЗОВАННЫЕ САМООТЧЕТЫ КАК МЕТОД ИЗУЧЕНИЯ ЛИЧНОСТИ

Хотя термин «личность» (*personality*) иногда употребляется в более широком смысле, согласно принятой в психометрии терминологии к «тестам личности» (*personality tests*) или, как чаще говорят, «личностным тестам» относятся инструменты для измерения характерных особенностей эмоциональной и мотивационной сферы, межличностных отношений и аттитюдов индивидуума, отличаемых таким образом от способностей.[1] В следующих четырех главах будут рассмотрены основные разновидности личностных тестов. С этой целью все доступные на данный момент инструменты классифицируются в соответствии с методом получения данных от обследуемых лиц. Эта глава посвящена личностным опросникам, относящимся к классу стандартизованных самоотчетов. В главе 14 будут рассмотрены методики для измерения интересов и аттитюдов. Тесты, охватываемые этими двумя главами, являются преимущественно бланковыми формами стандартизованного самоотчета, пригодными для группового обследования, хотя многие из них, конечно же, можно использовать в индивидуальном оценивании. Применение проективных методик для оценки характерных особенностей личности обсуждается в главе 15. В главе 16 дается обзор других, весьма разнородных подходов к оценке личности, частью находящихся еще на стадии экспериментальной разработки.

Число доступных пользователям личностных тестов достигает нескольких сотен. Особенно многочисленны личностные опросники и проективные методики. В этой книге мы будем рассматривать главным образом типы подходов к оценке личности. Каждый подход иллюстрируется кратким описанием некоторых из наиболее известных тестов, разработанных в его русле. Написано несколько книг, посвященных исключительно оценке личности посредством тестов, а также с помощью других методик. Для более подробного знакомства с данной темой, мы отсылаем читателя к специализированным изданиям (например, Angleitner, & Wiggins, 1986; Butcher, 1995; Lanyon & Goodstein, 1997; J. S. Wiggins, 1973/1988).

---

[1] Что касается более полного освещения современных подходов к теории и исследованиям личности, см. L. R. Aiken (1993), Burger (1993) и Maddi (1989).

При разработке личностных опросников различия в подходах проявляются в формулировке, компоновке, отборе и группировке вопросов. Среди используемых в настоящее время основных методик есть такие, в основу которых положены 1) значимость (релевантность) содержания (*content relevance*), 2) привязка к эмпирическому критерию, 3) результаты факторного анализа и 4) теория личности. В последующих разделах каждый из подходов будет рассмотрен и проиллюстрирован на примере конкретных тестов. Следует, однако, отметить, что эти подходы не являются альтернативными или взаимоисключающими. Теоретически их все можно объединить при разработке одного личностного опросника. На практике в большинстве нынешних опросников используются два и более из этих подходов.

Хотя некоторые личностные тесты используются в качестве средств группового отсеивания, большинство находит свое основное применение в условиях клиники и при консультировании. Поэтому следующие четыре главы хорошо бы читать с учетом перспектив этих специализированных контекстов тестирования, изложенных в главе 17. В своем настоящем виде большинство личностных тестов следует рассматривать либо как вспомогательные средства при индивидуальном оценивании, либо как исследовательские инструменты.

Методики, основанные на отборе релевантного содержания

Прототипом личностных опросников, относящихся к классу стандартизованных самоотчетов, послужил разработанный в годы Первой мировой войны «Бланк личных сведений» Вудвортса (см. ссылки в главе 2). Этот опросник, в сущности, представлял собой попытку стандартизовать психиатрическое интервью и приспособить эту процедуру для массового тестирования. В соответствии с поставленной задачей Вудвортс собрал сведения по общим невротическим и предневротическим симптомам, используя для этого литературу по психиатрии и консультации у психиатров. Первый вариант опросника как раз строился по принципу выявления подобных симптомов. Вопросы касались таких отклонений в поведении, как аномальные страхи или фобии, обсессии и компульсии, кошмары и другие расстройства сна, чрезмерная утомляемость и другие психосоматические симптомы, дереализация и двигательные расстройства, такие как тики и тремор. При окончательном отборе вопросов Вудвортс применил определенные виды статистической проверки, которые будут рассмотрены в следующем разделе. Тем не менее очевидно, что при построении и использовании этого опросника основной упор делался на релевантность содержания его вопросов. На это указывают как источники, откуда брались вопросы, так и общепризнанность некоторых типов поведения как неадаптивного. О значении наследия Вудвортса для конструирования современных личностных опросников говорит хотя бы тот факт, что при создании банка новых вопросов инструкции для разработчиков до сих пор основываются, в основном, на контент-анализе области оцениваемого поведения.

Современным примером содержательного подхода к разработке стандартизованных самоотчетов является Контрольный перечень симптомов-90, пересмотренная версия (*Symptom Checklist-90-Revised* [*SCL-90-R*] — Derogatis, 1994). *SCL-90-R* предназначен для скрининга психологических проблем и психопатологических симптомов.

Он состоит из кратких описаний 90 симптомов (например, Плохой аппетит; Слабость или головокружение). Респондентов просят указать, используя 5-балльную шкалу, сколь сильно они страдали от каждой из проблем в течение последних семи дней. *SCL-90-R* имеет отдельные нормы для взрослых мужчин и женщин, для подростков, не обращавшихся за психиатрической (или психологической) помощью, а также для психически больных лиц (как госпитализированных, так и живущих в домашних условиях). Однако некоторым из этих норм не достает репрезентативности; например, нормативная группа госпитализированных психически больных состояла преимущественно из лиц с низким социоэкономическим статусом, а подростковая группа была составлена почти целиком из представителей белого населения, относящегося в основном к среднему классу.

Пункты (или как еще говорят, вопросы, хотя это и не совсем корректно) *SCL-90-R* имеют сходство с пунктами более ранних инструментов такого типа не только в том, что они отбирались на основе принципов релевантности содержания и клинической полезности, но и в том, что некоторые из них восходят через промежуточные шкалы вроде Контрольного перечня симптомов Хопкинса (*Hopkins Symptom Checklist*) и Медицинский индекс Корнелла (*Cornell Medical Index*) к классическому Бланку личных сведений Вудвортса (Derogatis, & Lazarus, 1994). Эти пункты организованы в девять измерений (*dimensions*) психопатологии: Соматизация (*Somatization*), Депрессия, Тревожность, Враждебность, Психотизм, Межличностная чувствительность (*Interpersonal Sensitivity*), Фобическая тревога, Параноидная идеация (*Paranoid Ideation*), Обсессивно-компульсивные симптомы. Факторно-аналитические исследования этих шкал говорят о том, что они коррелируют друг с другом и поэтому не очень-то полезны в дифференциальной диагностике. Тем не менее общие индексы *SCL-90-R* оказались надежными индикаторами наличия и тяжести психопатологии (Payne, 1985). Этот Контрольный перечень и родственные ему инструменты, такие как Краткий инвентарь симптомов (*Brief Symptom Inventory*), целесообразно использовать как составную часть батареи при оценке изменений в ходе терапии и в исследованиях результатов различных курсов лечения.

Главное преимущество содержательного подхода к разработке личностных опросников заключено в его простоте и непосредственности. Хотя эти качества делают возможным получение относительно кратких и экономичных методик, их прозрачность предоставляет обследуемым больше возможностей для сознательной фальсификации результатов, чем другие методы. Основанные на релевантном содержании измерительные инструменты обычно не содержат защитных мер, предназначенных для предотвращения или обнаружения тенденциозности в ответах (Bornstein, Rossner, Hill, & Stepanian, 1994), — проблема, которую мы будем обсуждать позднее в этой главе. Поэтому не рекомендуется принимать решения, опираясь исключительно на результаты, полученные с помощью таких инструментов.

Привязка к эмпирическому критерию

Основной подход. *Привязка к эмпирическому критерию (empirical criterion keying)* касается разработки системы оценивания ответов на основе какого-то внешнего критерия. Этот метод заключается в отборе пунктов опросника, которые предполагается сохранить в его составе, и приписывание количественных весов каждому ответу. При

конструировании уже упоминавшегося Бланка личных сведений Вудвортса некоторые виды статистической проверки, примененные при окончательном отборе вопросов, подсказали способ привязки к критерию. Так, в этом опроснике исключался любой пункт, если 25 % или более испытуемых из нормальной выборки ответили на него в нежелательном направлении. Правомерность этой процедуры объяснялась тем, что столь часто встречающаяся поведенческая характеристика в выборке нормальных людей не может служить признаком анормальности. При отборе вопросов применялся также метод контрастных групп. В перечне симптомов сохранялись только те, которые встречались, по крайней мере, вдвое чаще в группе лиц, предварительно диагностированных как психоневротики, чем в группе нормальных людей.

Однако несмотря на частичное использование таких эмпирических проверок, содержательные подходы опираются на предположение о буквальном или соответствующем действительности понимании пунктов опросника. Ответ на каждый вопрос рассматривается как показатель фактического наличия или отсутствия конкретной проблемы, мнения или поведения, описанных в данном вопросе. С другой стороны, при привязке к эмпирическому критерию ответы трактуются как диагностические или симптоматические относительно критериального поведения, с которым выявлена их связь. Впервые излагая этот подход, Мил (Meehl, 1945) писал:

> ...*вряд ли* можно счесть самой плодотворной идею рассматривать вербальный тип личностного опросника как «самооценивание» или самоописание, ценность которого требует допущения о точности наблюдений тестируемого за самим собой. Вернее воспринимать ответ на вопрос теста как интересный сам по себе фрагмент вербального поведения, знание которого может иметь большее значение, чем любое знание «фактического» материала, на поверхностное выявление которого нацелен данный вопрос. Так, если ипохондрик говорит, что у него «постоянно болит голова», интересен сам факт, что он *говорит это* (р. 9).

Опросники типа самоотчетов, несомненно, представляют собой серии стандартизованных вербальных стимулов. Когда при создании опросников следуют методам привязки к эмпирическому критерию, ответы, вызываемые этими стимулами, оцениваются исходя из их эмпирически установленных коррелятов поведения. Иначе говоря, они трактуются подобно ответам на любые другие психологические тесты. То, что ответы опросника могут соответствовать *восприятию* (*perception*) субъектом реальности, дела не меняет, поскольку такое соответствие — только одна из гипотез, объясняющая эмпирически устанавливаемую валидность некоторых пунктов опросника.

Миннесотские многофазные личностные опросники

Самым известным примером использования метода привязки к эмпирическому критерию при конструировании личностного теста является Миннесотский многофазный личностный опросник (*Minnesota Multiphasic Personality Inventory* [*MMPI*]). В последние годы *MMPI* был подвергнут переработке, результатом которой стало создание двух отдельных версий: *MMPI-2* (Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989) и *MMPI-Adolescent* — Подростковая версия (*MMPI-A* — Butcher et al., 1992). Несмотря на существование этих более новых версий, невозможно обсуждать ни одну из них без обращения к оригинальному *MMPI* и без учета той роли, какую он сыграл в истории оценки личности. Хотя в рамках этого учебника и не ставится цель

дать сколько-нибудь подробную характеристику *MMPI*, следует все же отметить, что этот инструмент был самым используемым и самым исследованным личностным тестом.[1]

Во многих отношениях MMPI — как измерительный инструмент — стал жертвой собственного успеха. Идея создания подобного инструмента возникла в 1930-х гг., у клинического психолога С. Р. Хатуэя (Starke R. Hathaway) и нейропсихиатра Дж. Ч. Мак-Кинли (J. Charnley McKinley), которые первоначально опубликовали этот опросник в серии статей в 1940-х гг., представив как вспомогательное средство для постановки психиатрического диагноза.[2] Впоследствии его эффективность в том, что касается обнаружения психопатологии и дифференциальной диагностики на основе тогда еще очень грубых нозологических категорий, привела к неуклонно расширяющемуся использованию *MMPI* в целях, явно расходившихся с его первоначальным предназначением. К 1960-м гг. *MMPI* занял прочное положение *главного* личностного теста и применялся не менее часто, а может быть и чаще, для тестирования нормальных людей в ходе консультирования, приема на работу и на военную службу, медицинского обследования и судебно-медицинской экспертизы, чем для оценки психически больных. К 1980-м гг. литература по *MMPI* насчитывала несколько тысяч ссылок, документально подтверждающих, помимо многих других вещей, использование его 13 основных шкал в работе с большим количеством разнообразных популяций, создание на основе его вопросов сотен специальных шкал и внушительную массу эмпирических коррелятов высоких оценок по отдельным шкалам и различных паттернов профиля. Однако примерно в это же время документально зафиксированные концептуальные и психометрические проблемы *MMPI* стали вызвать большее беспокойство у его пользователей в свете определенных достижений в психопатологии и теории личности, а также в области конструирования тестов. Более того, к этому времени стало совершенно ясно, что построенные на узком базисе и теперь уже устаревшие нормы по данному тесту не годятся для нынешних испытуемых, и поэтому сам основополагающий принцип определения анормальности опирался на ненадежный эмпирический фундамент (Colligan, Osborne, Swenson, & Offord, 1983, 1989). Фактически, первоначальная выборка стандартизации оказалась в чем-то похожей на ненормативную эталонную группу, исходя из результатов которой определялась шкала показателей. И только собранные позже более обширные данные, касающиеся кодов профиля (*profile codes*), обеспечили основу для нормативной интерпретации.

Таким образом, комитет, которому была поручена рестандартизация *MMPI*, столкнулся с нелегкой задачей модернизации этого инструмента, пытаясь сохранить все богатство интерпретативного материала, относящегося к оценке личности и, особенно, психопатологии, вложенного в базисную структуру *MMPI*. Ради сохранения преемственности этот комитет решил оставить подавляющее большинство пунктов всех оригинальных клинических шкал и шкал валидности, а также многие из дополнительных шкал, со всеми присущими им недостатками. К важнейшим изменениям относятся: 1) полное обновление норм опросника; 2) разработка равномерных Т-показателей

[1] Довольно краткое описание оригинального *MMPI* можно найти в более ранних изданиях этого учебника (например, Anastasi, 1988). Более исчерпывающие трактовки этого инструмента даны в томах классического справочного руководства по *MMPI* (Dahlstrom, Welsh, & Dahlstrom, 1972, 1975).

[2] Эти первые статьи перепечатаны в хрестоматии под ред. Далстромов (Dahlstrom, & Dahlstrom, 1988).

для восьми оригинальных клинических шкал и для всех шкал содержания (*content scales*); 3) пересмотр и устранение устаревших или вызывающих иные возражения пунктов и добавление новых; 4) создание нескольких новых шкал валидности, дополнительных шкал и шкал содержания; 5) разделение опросника на две версии, предназначенные для разных возрастных групп.

Миннесотский многофазный личностный опросник–2. *MMPI-2* состоит из 567 пунктов в виде утвердительных высказываний, на которые тестируемый дает ответ «верно» или «неверно». Первые 370 пунктов, которые, в сущности, идентичны пунктам *MMPI*, если не считать редакционных правок и иного расположения, обеспечивают получение всех ответов, необходимых для расчета показателей по сохранившимся из оригинальной версии опросника 10 «клиническим» шкалам и 3 шкалам «валидности». Оставшиеся 197 пунктов (107 из которых совершенно новые) необходимы для вычисления показателей по всему комплекту из 104 новых, пересмотренных и сохраненных шкал валидности, шкал содержания и дополнительных шкал и подшкал, входящих в состав полного опросника. Представленные в пунктах опросника утверждения широко варьируют по содержанию, охватывая такие области, как 1) общее состояние здоровья, 2) эмоциональные, неврологические и двигательные симптомы, 3) сексуальные, политические и социальные аттитюды, 4) учебные, профессиональные, семейные и супружеские проблемы, а также 5) многие известные невротические или психотические поведенческие проявления, включая обсессивные и компульсивные состояния, мании, галлюцинации, идеи отношения, фобии, садистические и мазохистские наклонности. Далстром (Dahlstrom, 1993a) подготовил приложение к руководству, в котором дана вся информация, необходимая для сравнения пунктов *MMPI-2* с пунктами классического *MMPI*. Ниже приведено несколько примеров утверждений вместе с номерами пунктов, в которых они помещены в современной форме теста[1]:

> У меня беспокойный, прерывистый сон. (39)
> Я уверен, что против меня что-то замышляется. (138)
> Меня беспокоят сексуальные проблемы. (166)
> Когда мне скучно, я стараюсь как-то себя возбудить. (169)
> Большинству людей в душе не нравится причинять себе неудобства, помогая другим. (286)

MMPI-2 обеспечивает получение показателей по 10 базисным «клиническим» шкалам, которые являются теми же самыми, что и в оригинальной версии *MMPI*:

1. *Hs*: Ипохондрия
2. *D*: Депрессия
3. *Hy*: Истерия
4. *Pd*: Психопатическое отклонение
5. *Mf*: Маскулинность-феминность
6. *Pa*: Паранойя
7. *Pt*: Психастения
8. *Sc*: Шизофрения

[1] Из: Minnesota Multiphasic Personality Inventory–2. Copyright © by The Regents of the University of Minnesota. All rights reserved. Воспроизводится с разрешения.

9. *Ma*: Мания
10. *Si*: Социальная интроверсия

Восемь из этих шкал были построены в 1940-х гг. эмпирически, путем привязки их пунктов к критериям, по которым небольшие клинические выборки, большей частью около 50 человек каждая, с типичными для того времени психиатрическими диагнозами, отличались от нормальной контрольной группы, состоявшей из 724 посетителей и родственников пациентов клиники Миннесотского университета (Hathaway, & McKinley, 1940, 1943). Шкала маскулинности—фемининности, первоначально предназначавшаяся для различения мужчин с гомосексуальной или гетеросексуальной ориентацией, разрабатывалась исходя из различий в частоте ответов на конкретные пункты у военнослужащих-мужчин и женского персонала авиакомпаний. Показатели по этой шкале свидетельствуют о степени соответствия интересов и аттитюдов индивидуума стереотипам его половой группы. Добавленная позднее шкала социальной интроверсии была получена в результате анализа ответов двух контрастных групп студентов колледжа, отобранных на основе крайних значений показателей по тесту интроверсии—экстраверсии.

Оставляя базисные шкалы *MMPI* практически нетронутыми (не считая девяти удаленных пунктов, нескольких редакторских правок и изменения порядка расположения пунктов), разработчики *MMPI-2* стремились сохранить богатство клинически полезной информации, связанной с интерпретацией кодов профиля (*profile codes*), основанных на паттернах показателей по этим шкалам (Graham, 1993; Greene, 1991). Однако при этом сохранились и присущие этим шкалам устарелые понятия психопатологии и все последствия наивного и не вполне корректного применения эмпирического метода привязки к контрастным критериальным группам. Остались нерешенными и другие проблемы, такие как многомерность и частичное перекрытие базисных шкал (Helmes, & Reddon, 1993).

Отличительной особенностью оригинального *MMPI* было использование в нем трех так называемых шкал валидности, которые были сохранены и в *MMPI-2*.[1] Эти шкалы не имеют отношения к валидности в специальном смысле этого слова. В сущности, они представляют собой средства контроля небрежности, непонимания, симуляции, а также действия аттитюдов тестируемого и его установок на определенный тип ответов. Показатели валидности включают:

Показатель лжи (Lie Score), или L-показатель: основан на группе утверждений, содержание которых провоцируют тестируемого представить себя в благоприятном свете, но одновременно сводит к минимуму вероятность того, что правдивый человек будет отвечать на них именно в такой манере (например, «Мне не нравится никто из тех, кого я знаю»).

Показатель редкости (Infrequency Score), или F-показатель: определяется по набору из 60 (в оригинальной версии — из 64) утверждений, ответы на которые в засчитываемом направлении дало не более 10 % группы стандартизации *MMPI*. Хотя в этих утверждениях изображается нежелательное поведение, они не входят в какой-то специфический паттерн анормальности. Следовательно, малове-

[1] Есть еще показатель «не могу сказать» (?), отображающий число пунктов, которые были двояко отмечены в листе для ответов или вообще пропущены. Если этот показатель превышает 30 пунктов, результаты теста считаются весьма сомнительными и, вероятно, недействительными.

роятно, что любой отдельный человек обнаружит все или почти все из этих симптомов. Высокий *F*-показатель может указывать на ошибки в подсчете баллов, небрежность в ответах, сильную эксцентричность, психотические процессы или на преднамеренную симуляцию.

Показатель коррекции (Correction Score), или K-показатель: благодаря использованию еще одной комбинации специально подобранных пунктов, этот показатель служит, предположительно, более тонкой мерой аттитюда в отношении заполнения данного опросника. Высокий *K*-показатель может указывать на оборонительную позицию или на попытку «прикинуться хорошим». Низкий *K*-показатель может свидетельствовать о чрезмерной откровенности и самокритичности или о нарочитой попытке «показаться плохим».

Первые два показателя *(L и F)* первоначально использовались для общей оценки ответов, зафиксированных в протоколах тестирования. Если любой из этих показателей превышает заданное значение, то результаты теста считаются недействительными. С другой стороны, *K*-показателю отводилась роль переменной-подавителя (*suppressor variable*). Он используется для вычисления поправочного коэффициента, прибавляемого к показателям некоторых клинических шкал для получения уточненных суммарных показателей. В силу сомнительной эффективности такого использования *K*-показателя основные показатели по «неустойчивым» клиническим шкалам могут приводиться как с этой поправкой, так и без нее. Хотя необычайно высокое значение *K*-показателя само по себе может вызвать недоверие к результатам теста и послужить основанием для дополнительного тщательного обследования, умеренные превышения по шкале *K* могут на самом деле отражать силу эго и позитивный настой испытуемого. Особенно важно оценивать такие превышения в свете истории и жизненных обстоятельств конкретного человека.

Среди дополнительных шкал *MMPI-2* (а их 21) имеются три новых индикатора «валидности», которые помогают оценить внимательность и правдивость отвечающих на опросник лиц. Это — Вспомогательная *F*-шкала (*Back F* [сокращенно F_b]), Шкала несовместимости изменяемых ответов (*Variable Response Inconsistency Scale* [сокращенно *VRIN*]) и Шкала несовместимости правдивых ответов (*True Response Inconsistency Scale* [сокращенно *TRIN*]). Тогда как F_b-шкала является по существу расширением оригинальной *F*-шкалы для утверждений, с которыми тестируемый сталкивается во второй половине опросника, *VRIN* и *TRIN* — это новые шкалы, состоящие из пар утверждений со сходным или противоположным смыслом и нацеленные на обнаружение непоследовательных или противоречивых ответов.

Основная форма профиля *MMPI-2* (рис. 13–1) включает 13 шкал (валидности и клинических), перенесенных из оригинальной версии. Кроме того, имеются отдельные формы профилей для 15 шкал содержания (*content scales*), 27 составляющих шкал содержания (*content component scales*), 21 дополнительной шкалы (*supplementary scales*) и 28 подшкал Харриса–Лингоуза (*Harris–Lingoes subscales*).[1] Некоторые из этих шкал и подшкал совершенно новые, а некоторые были оставлены из оригинала, но все они дают показатели, рассчитываемые на основе нормативной выборки *MMPI-2*, состоящей из 2600 взрослых в возрасте от 16 до 84 лет. Эта выборка является гораздо

---

[1] С 1996 г. подшкалы искусности–безыскусности Винера–Хармона (*Wiener–Harmon Subtle-Obvious subscales*) уже нельзя получить от издателей *MMPI-2*.

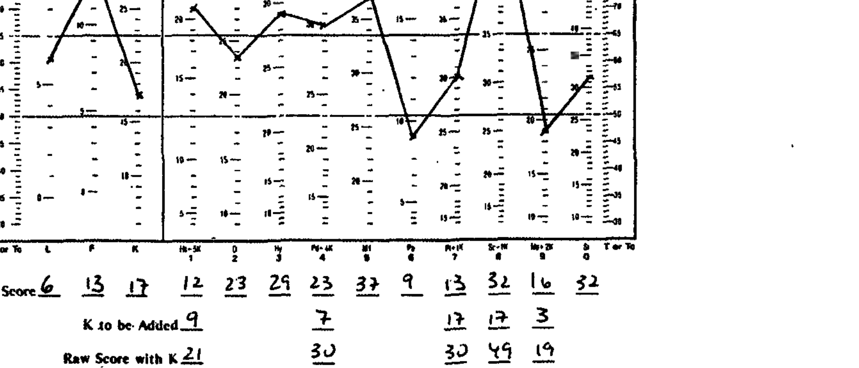


Рис. 13-1. Профиль основных шкал *MMPI-2* 41-летнего мужчины. Код в обобщенной форме фиксирует отношения между показателями шкал *MMPI-2* и обеспечивает основу для интерпретации профиля и различных сравнений. *(Значения показателей в данной иллюстрации профиля позаимствованы из* Annotated Sample Report *для* The Minnesota Report Adult Clinical System – Revised *[University of Minnesota, 1993, p. 2–7]).*

более репрезентативной по отношению к современному американскому населению, чем исходная миннесотская нормативная группа, и комплектовалась в семи разных штатах в стремлении отразить население США с учетом важнейших демографических переменных, включая пол, возраст и этническую принадлежность (Dahlstrom & Tellegen, 1993). И все же репрезентативность нормативной выборки *MMPI-2* была подвергнута сомнению, в первую очередь из-за ее высокого образовательного и профессионального уровней и непропорционально низкой представленности выходцев из Азии и испаноязычных американцев, если судить по данным переписи 1980 г. (Duckworth, 1991).

Главным результатом пересмотра норм *MMPI* стало снижение «демаркационной линии» клинических профилей. Это изменение, о котором многие предупреждали, вероятно, обусловлено поколенными различиями, действием особых факторов исключительно в первичной миннесотской выборке и тем, как она использовалась при разработке *MMPI* (D. S. Nichols, 1992). Во всяком случае, критический *T*-показатель, необходимый для того, чтобы считать подъем профиля по какой-либо шкале представляющим клинический интерес, равен теперь 65, или примерно на 1,5 *SD* выше среднего, против прежних 70. Еще одно нововведение в *MMPI-2* состоит в использовании равномерных — в противоположность получаемым путем линейного преобразования или нормирования — *T*-показателей в 8 из 10 клинических шкал и во всех шкалах содержания. Оно предполагает выравнивание показателей по всем этим шкалам относительно среднего смешанного распределения. Равномерные *T*-показатели допускают сравнения между шкалами в единицах процентильных эквивалентов, не приводя к сколько-нибудь заметному искажению формы распределений первичных оценок, обладающих все как одно положительной асимметрией, хотя и выраженной в различной степени (Tellegen, & Ben-Porath, 1992).

Только что описанные изменения безусловно оправданы с точки зрения психометрических требований. Однако из-за того что эти изменения имели следствием появление различий между паттернами профилей (и кодов) *MMPI* и *MMPI-2*, развернулась широкая полемика по поводу законности распространения сведений из обширной литературы по вопросам интерпретации профиля *MMPI* и из многолетнего опыта применения этого опросника в клинике на результаты, получаемые с помощью *MMPI-2* (Chojnacki, & Walsh, 1992; Morrison, Edwards, & Weissman, 1994; Tellegen & Ben-Porath, 1993). Имеющиеся данные говорят в пользу того, что для четких профилей с ясно выраженным расхождением шкальных показателей, по-видимому, имеется почти такое же соответствие между типами кодов *MMPI* и *MMPI-2*, как и между типами кодов любой из этих версий, полученными при первом и повторном предъявлении опросника (Archer, 1992b; Graham, 1993). Во всяком случае, руководство по *MMPI-2* содержит информацию, позволяющую пользователям сравнить показатели, выводимые из этих двух версий, на основе ответов на любую из них. И хотя это предложение создает свои проблемы (см., например, Ben-Porath, & Tellegen, 1995), оно было поддержано некоторыми специалистами как эмпирически оправданный метод, помогающий пользователям преодолеть период перехода от одной версии опросника к другой (см., например, Humphrey, & Dahlstrom, 1995).

Миннесотский многофазный личностный опросник — Подростковая версия. *MMPI-A* — новая форма *MMPI*, разработанная специально для работы с подростками. Подростковая версия вобрала в себя большинство качеств *MMPI* и *MMPI-2*, включая

все 13 базисных шкал, но более приспособлена к юному возрасту тестируемых за счет 1) сокращения длины опросника до 478 пунктов, 2) включения новых пунктов и шкал, охватывающих существенно значимые именно для них проблемные области — школу и семью, и, главное, 3) наличия норм, соответствующих возрасту. При разработке *MMPI-A* использовалась нормативная выборка, сформированная из 1620 современных подростков в возрасте от 14 до 18 лет; одновременно с ней для проведения сравнений и исследований валидности была набрана клиническая выборка — 713 подростков того же возраста.

В добавление к базисным клиническим шкалам и шкалам валидности, общим с *MMPI-2*, *MMPI-A* имеет свои собственные шкалы валидности (*F1* и *F2*); то же самое касается его шкал содержания и дополнительных шкал и подшкал: некоторые из них являются общими со шкалами и подшкалами опросника для взрослых, а некоторые существуют только в подростковой версии. Хотя внушительные исследования (Dahlstrom et al., 1972; Marks, Seeman, & Haller, 1974), включая опубликованные Далстромом с соавторами нормы и таблицы пересчета показателей, подтвердили правомерность использования *MMPI* в работе с подростками, полученные в них результаты совсем не обязательно применимы к *MMPI-A*, который является скорее инструментом нового типа, чем новой редакцией. Полезность *MMPI-A* как такового можно будет определить только после накопления исследовательских данных и материалов интерпретации, начавшегося одновременно с его публикацией (Archer, 1992a; Butcher, & Williams, 1992; Williams, Butcher, Ben-Porath, & Graham, 1992).

Заключительные комментарии по поводу Миннесотских многофазных личностных опросников. Вопреки своему происхождению, которому *MMPI* обязан славой прототипичного продукта наивного эмпиризма, и несмотря на непрекращающиеся слухи о его близкой кончине, ему удалось уцелеть и остаться популярным рабочим инструментом. Фактически, хотя его первоначальное предназначение состояло в том, чтобы служить вспомогательным средством психиатрической диагностики, а методы, используемые при его разработке, делали его непригодным для оценки личности нормальных людей, *MMPI* широко использовался в работе с нормальными, а не только с психически больными людьми. *MMPI* и его пересмотренные версии включили в свою первоначальную, эмпирически полученную основу ряд дополнительных процедур и стратегий интерпретации. Те качества, о которых уже упоминалось мимоходом, свойственны многим шкалам, разработанным путем группирования пунктов опросника на основе их содержания (Butcher, Graham, Williams, & Ben-Porath, 1990), и это стоит особо отметить, так же как и использование факторного анализа при разработке некоторых из дополнительных шкал (Welsh, 1956).

Продолжают развиваться новые подходы к решению сложной задачи интерпретации результатов *MMPI*. Один из самых современных состоит в использовании структурных сводок (*structural summaries*), которые придают хоть какую-то связность многообразию взаимно коррелированных шкал Миннесотских опросников и облегчают пользование ими. Этот подход, основанный на анализе шкал, подшкал и отдельных пунктов, имеет целью сокращение числа измерений (*dimensions*), необходимых для интерпретации результатов тестирования, за счет отбрасывания таких произвольных классификаций, как «дополнительные» шкалы, шкалы «содержания» и шкалы «Харриса—Лингоуза». Наиболее выраженные измерения (*dimensions*) используются для организации категорий в формате «Структурной сводки» (*Structural Summary*), по-

добном разработанному Экснером (*Exner*) для работы с тестом Роршаха (см. главу 15). Этот подход к интерпретации *MMPI* находится на начальных стадиях развития и нуждается в дополнительном исследовании и обосновании. Тем не менее он уже применяется в работе как с *MMPI-2*, применительно к которому структурные измерения (*structural dimensions*) установлены, в основном, на основе контент-анализа, так и с *MMPI-A*, в котором используются измерения, полученные посредством факторного анализа, причем в обоих случаях результаты выглядят многообещающими (Archer, & Krishnamurthy, 1994; Archer, Krishnamurthy, & Jacobson, 1994; Nichols, & Greene, 1995).

Два других направления развития, в которых *MMPI-2* и *MMPI-A* претерпели быстрые изменения, впрочем, как и большинство других тестов, — это, во-первых, компьютеризация процедур предъявления заданий, подсчета показателей и интерпретации результатов и, во-вторых, рост и совершенствование переводов этих инструментов на многие языки. Традиционные компьютерные программы предъявления опросников и подсчета показателей, которые существовали для *MMPI*, доступны и для новых версий этих опросников, как и ряд услуг автоматизированной интерпретации. В добавление к этому была разработана и успешно опробована компьютеризованная адаптивная версия *MMPI-2* (Roper, Ben-Porath, & Butcher, 1991, 1995).

В то время как потребовалось почти десять лет, чтобы сделать первый перевод оригинального *MMPI*, работа по кросс-культурной адаптации *MMPI-2* была начата еще до его публикации. За первые три года его существования появилось 15 проектов перевода, часть из которых завершена, а некоторые еще выполняются. Было подготовлено шесть переводов (или адаптаций) *MMPI-2* на испанский язык, в том числе две версии для работы с испаноязычными американцами. Имеется также справочник по переводам и международным адаптациям *MMPI-2* (Butcher, 1996). Один примечательный вывод тех, кто собирал данные с использованием таких переводов и адаптаций, состоит в том, что показатели современных кросс-культурных выборок нормальных лиц ближе к нормам стандартизации *MMPI-2*, чем показатели прежних аналогичных выборок к нормам *MMPI*.

Пытаясь усовершенствовать этот классический образец личностного опросника, не изменяя его ни в одном принципиальном отношении, Комитет по рестандартизации *MMPI* поставил перед собой две трудные и в значительной степени противоречащие друг другу цели. Время покажет, помогут ли принятые комитетом решения сохранить Миннесотским опросникам свое превосходство в новом столетии, или же семейство *MMPI* уступит лидерство новому поколению аналогичных инструментов, таких как Базисный личностный опросник Джексона (*Basic Personality Inventory*), обсуждаемый в этой главе несколько позже, или Опросник для оценки личности (*Personality Assessment Inventory* [*PAI*]), разработанный Лесли Мори (Leslie Morey) с использованием сложной стратегии, сочетающей логические и эмпирические методы в целях обеспечения психометрической доброкачественности его шкал. Между тем пока явно преждевременно ожидать снижение темпов издания и ассортимента книг и статей по *MMPI-2* и *MMPI-A* (Butcher, 1990; Butcher, Graham, & Ben-Porath, 1995; Keller & Butcher, 1991; Pope, Butcher, & Seelen, 1993).

Калифорнийский психологический опросник

На протяжении всего своего существования *MMPI* служил основой для разработки других широко используемых опросников, среди которых самым известным явля-

ется Калифорнийский психологический опросник (*California Psychological Inventory* [*CPI*]). Несмотря на то что около половины пунктов *CPI* было позаимствовано из *MMPI*, этот опросник был создан специально для работы с нормальными взрослыми популяциями. В своей самой свежей редакции — третьей по счету — *CPI* состоит из 434 пунктов, на которые требуется ответить «верно» или «неверно», и дает показатели по 20 шкалам (Gough, & Bradley, 1996). Три из них являются шкалами «валидности», предназначенными для оценки аттитюдов испытуемых в отношении тестирования. Эти шкалы обозначаются как: *Благополучие (Well-being, или сокращенно Wb)* — основывается на ответах нормальных людей, которых просили «притвориться плохими»; *Благоприятное впечатление (Good impression, или сокращенно Gi)* — основывается на ответах нормальных людей, которых просили «прикинуться хорошими», и *Общность (Communality, или сокращенно Cm)* — основывается на подсчете частоты наиболее популярных ответов. Остальные 17 шкал дают показатели по таким измерениям личности, как доминирование, общительность, самопринятие, ответственность, социализация, самоконтроль, достижение-через-подчинение, достижение-через-независимость, эмпатия и независимость. Последние две шкалы были добавлены при пересмотре опросника в 1987 г.

Для 13 из этих 17 шкал пункты-утверждения отбирались на основе ответов контрастных групп, противопоставляемых относительно таких критериев, как школьные отметки, принадлежность к социальному классу, участие в факультативных занятиях и рейтингов (субъективных оценок). Эти рейтинги были получены посредством номинаций лиц своего круга (*peer nominations*), что оказалось эффективной методикой оценки многих межличностных черт (см. главу 16). Пункты-утверждения для остальных 4 шкал сначала группировались субъективно и затем проверялись на внутреннюю согласованность. Кросс-валидизация всех шкал на относительно больших выборках показала значимые различия между группами, хотя частичное перекрытие результатов контрастных по критерию групп было существенным, а корреляции с критерием — часто низкими.

Как и в *MMPI-2*, в *CPI* все результаты сообщаются в единицах шкалы стандартных показателей со средним, равным 50, и *SD* = 10; в настоящее время эта шкала базируется на нормативной выборке, состоящей из 3000 мужчин и 3000 женщин, извлеченной из архивов *CPI* таким образом, чтобы адекватно представлять генеральную совокупность — население США — с учетом возраста, социоэкономического уровня и географической зоны проживания. Предусмотрены нормы отдельно для женщин, отдельно для мужчин, а также для объединенной по полу выборки. В добавление к этому приводятся средние значения и стандартные отклонения показателей по каждой шкале для многих особых групп.

CPI, впервые опубликованный в 1956 г., был представлен на концептуальном уровне как «открытая система», элементы которой при необходимости можно удалять или, наоборот, добавлять (Gough, 1987, p. 1). В соответствии с этим представлением, при пересмотрах опросника он был сокращен от своей первоначальной длины в 480 пунктов сначала до 462 (пересмотр 1987 г.), а затем, в ходе последнего пересмотра, до 434 пунктов. Преследовалась цель исключить пункты, которые могли оказаться «неудобными» для некоторых респондентов или вызвать юридические возражения в свете Закона об инвалидах-американцах от 1990 г. (P. L. 101–336), особенно в ситуациях отбора персонала. Используя обширные архивные данные, накопленные более чем по 13 000 испытуемым, ответившим на все пункты *CPI*, Гау (Gough) и Брэдли (Bradley)

постарались сохранить надежность и валидность основных шкал путем поддержания длины шкал постоянной за счет замены исключенных пунктов другими, функционально эквивалентными с точки зрения их корреляции с критерием принадлежности к шкале.

Исследования с *CPI* дали большой объем информации, которая полезна при анализе профилей, строящемся как на превышениях показателей по отдельным шкалам, так и на конфигурациях или паттернах показателей по двум и более шкалам, в манере, близкой традиционному типу интерпретации кодов *MMPI* (McAllister, 1996). Кросскультурные исследования свидетельствуют о том, что *CPI* также полезен при изучении различий между этническими группами по измерениям личности (см. Dana, 1993; Davis, Hoffman, & Nelson, 1990). В добавление к основным шкалам было разработано — с помощью разных методов — еще несколько шкал *CPI*, примерами которых могут служить Шкала управленческого потенциала (*Managerial Potential Scale*) и Шкала трудовой направленности (*Work Orientation Scale*) (Gough, 1984, 1985). Подготовлено и руководство для профессиональных пользователей по применению *CPI* при отборе персонала и принятии решений о продвижении по службе (Meyer, & Davis, 1992).

Начиная с пересмотра 1987 г., в состав *CPI* включена также трехмерная типологическая модель для классифицирования индивидуумов с высокими и низкими показателями по трем структурным, или векторным, шкалам, идентифицированным методами факторного анализа и анализа заданий. Эти структурные шкалы измеряют такие факторы высшего порядка, как Интернальность/экстернальность, Принятие норм/отвергание норм и Самореализация. Показатели по первым двум шкалам используются для распределения респондентов по четырем типам личности (Альфа, Бета, Гамма и Дельта), тогда как показатели по третьей векторной шкале специально предназначены для оценки уровня интеграции или реализации положительного потенциала, связанного с типом человека. Типологическая модель несомненно является привлекательным элементом *CPI* для пользователей в области организационной психологии, однако эта модель была подвергнута серьезной критике за нечеткое описание того, каким образом ее получили (Engelhard, 1992). Вдобавок ко всему, она разделяет концептуальные и психометрические недостатки, присущие классификации людей на типы, основанной на произвольной дихотомизации одной или нескольких непрерывных переменных (*dimensions*).[1]

Личностный опросник для детей

Хотя в Личностном опроснике для детей (*Personality Inventory for Children* [*PIC*]) не использовались формулировки пунктов или данные *MMPI*, этот инструмент сконструирован на основе той же общей методологии, что и *MMPI* с *CPI* (Wirt, & Lachar, 1981; Wirt, Lachar, Klinedinst, & Seat, 1991). *PIC* разрабатывался в ходе 20-летнего исследования, проводимого группой специалистов из Миннесотского университета, которые находились в плену идеологии и опыта клинического применения *MMPI*. *PIC* предназначен для детей и подростков и охватывает возрастной диапазон от 3 до 16 лет. Главное различие между PIC и MMPI касается способа получения информации — ответы «верно»/«неверно» на пункты опросника дает не сам ребенок, а хорошо осведомленный взрослый, чаще всего мать. Такая процедура согласуется с общепринятой

[1] См. обсуждение Индикатора типов Майерса—Бригса в главе 16.

в детских клиниках практикой интервьюирования родителей, рассматриваемого в качестве главного источника информации о существующих проблемах и истории болезни. Этот опросник фактически обеспечивает систематический способ сбора такой информации и ее интерпретации исходя из нормативных и диагностических данных.

Оригинальный *PIC* содержал в сумме 600 пунктов, сгруппированных в 3 «шкалы валидности», шкалу общего скрининга и 12 клинических шкал. Шкалы валидности включают: 1) шкалу лжи (*Lie scale*), состоящую из пунктов, формулировки которых побуждают ребенка показать себя в нереалистично благоприятном свете; 2) шкалу частоты (*Frequency scale*), содержащую редко одобряемые пункты, и 3) шкалу дефензивности (*Defensiveness scale*), предназначенную для оценки защищающей родительской позиции в отношении поведения своего ребенка. Шкала скрининга (или отбора), а именно «Приспособление» (*Adjustment*), используется для выявления детей, нуждающихся в психологическом оценивании. Двенадцать клинических шкал разрабатывались для оценки когнитивного развития и учебных достижений, ряда хорошо определенных типов эмоциональных и межличностных проблем (например, депрессии, тревоги, ухода и гиперактивности), а также психологического климата в семье.

В современном формате пересмотренной версии *PIC* (*PIC-R*) изменен порядок и число пунктов, которых стало значительно меньше — 420. В буклетах для группового предъявления опросника они разбиты на три последовательно увеличивающиеся части. Часть I (пункты 1–131) дает показатели по шкале лжи и четырем новым, основанным на широких факторах шкалам. Часть II (пункты 132–280) *добавляет* к I части сокращенные версии других основных шкал и частичный набор критических пунктов. Часть III (пункты 281–420) прибавляет остальные пункты, необходимые для получения полного комплекта показателей по 16 первичным шкалам и 4 широким факторам и полного набора критических пунктов.

Семь из 16 шкал оригинального *PIC* разработаны путем эмпирических сравнений частоты ответов в критериальной и контрольной группах; благодаря итеративной процедуре пункты добавлялись на каждом этапе сравнений до достижения оптимальной валидности шкалы. При разработке других 9 шкал по существу использовались методы содержательной валидизации, в силу чего пункты первоначально отбирались на основе номинативных или ранговых оценок экспертов. Однако и в этих случаях прибегали к оценке внутренней согласованности ответов в рамках каждой шкалы и факторному анализу пунктов в целях повышения конструктной валидности шкал.

Составленное Лэйчером и Гдовски монографическое руководство (Lachar, & Gdowski, 1979) предлагает обширные интерпретативные данные для оригинального *PIC*, основанные на систематическом комплексном изучении валидности этого опросника. Дополнительное руководство, подготовленное для пересмотренной версии *PIC*, содержит столь же обширные данные по факторным шкалам и психометрические данные по сокращенным шкалам (Lachar, 1982). В добавление к этому Лэйчер и его коллеги использовали кластерный анализ для классификации разнородных выборок детей на основе их профилей *PIC* и изучили диагностически значимые характеристики, связанные с различными типами таких профилей. На базе этого продолжительного исследования были разработаны последовательные правила классификации профилей по типам и процедуры подсчета индексов сходства профилей (Gdowski, Lachar, & Kline, 1985; Kline, Lachar, & Gdowski, 1992). Актуарный[1] подход к интерпретации,

[1] То есть экспертно-статистический. — *Примеч. науч. ред.*

который использовался также в *MMPI* и *CPI*, представляет собой дальнейшее развитие и распространение традиционного способа привязки к эмпирическому критерию с этапа разработки теста на этап его интерпретации (см., например, Kline, Lachar, & Boersma, 1993).

Заслуживает внимания то обстоятельство, что *PIC-R* — это не форма стандартизованного самоотчета, а инвентарь наблюдаемого поведения. Как таковой, он созвучен с поведенческой оценкой (*behavioral assessment*), на которую ориентирована клиническая психология (см. главу 17). Однако получаемые от родителей сведения не имеют тех ограничений, которые были, в целом, признаны и отмечены авторами *PIC*. Как они указывают, ответы могут отчасти отражать мотивацию, аттитюды, личные эталоны и культурные нормы конкретного родителя. Поэтому можно ожидать некоторых несоответствий между отчетами различных наблюдателей, скажем двух родителей, и между родительскими отчетами и самоотчетами их детей, что, фактически, и обнаруживается. Один из способов разрешения этой неизбежной проблемы различающихся перспектив — попытаться оценить индивидуальные тенденции ответов, которые могли привести к искажению данных, с помощью «шкал валидности». Другая альтернатива — собрать и сравнить отчеты от нескольких наблюдателей. Правда, есть и третий путь — собрать и сравнить данные самоотчета и отчета наблюдателя.

Именно с этой целью был разработан Личностный опросник для юношества (*Personality Inventory for Youth* [*PIY*]) как средство стандартизованного самоотчета, сопоставимое с *PIC-R*. Хотя *PIY* может применяться самостоятельно, идеально, когда он обеспечивает основу для более комплексного оценивания, опирающегося на совместный анализ профилей, построенных по данным родителей и детей (Lachar, & Gruber, 1995a, 1995b). Авторы PIY использовали первые 280 пунктов *PIC-R* в качестве отправной точки при создании банка формулировок, большинство которых было преобразовано из формы третьего в форму первого лица (например, формулировка «Мой ребенок *часто приводит друзей в дом» заменялась на «Я часто привожу друзей к себе домой»*). Формулировки некоторых пунктов пришлось подвергнуть более существенной редакции, чтобы привести в соответствие с заданным значением, а от некоторых пунктов пришлось и вовсе отказаться как неподходящих для формы самоотчета на выбранном возрастном уровне (в интервале от 9 до 18 лет) вследствие понятных трудностей у младших испытуемых. Было включено и несколько совершенно новых пунктов. Окончательная версия *PIY* состоит из 270 пунктов, составляющих 9 не перекрывающихся клинических шкал, 24 не перекрывающиеся подшкалы и 4 «шкалы валидности». Первые 80 пунктов могут использоваться как сокращенная форма *PIY* для целей скрининга. *PIY* разработан и стандартизован с использованием выборки 2337 учащихся нормальных школ и клинической выборки 1178 детей и подростков (Lachar & Gruber, 1993).

Еще многое предстоит сделать в отношении как *PIC-R*, так и *PIY*. Следует обновить нормы PIC, основанные на данных, собранных в конце 1950-х — начале 1960-х гг., а также расширить его исследовательскую базу, с тем чтобы охватить дошкольный возраст (Knoff, 1989). Что касается *PIY*, еще предстоит доказать его клиническую полезность как самостоятельного инструмента и в сочетании с *PIC-R*. Тем не менее оба эти инструмента имеют под собой впечатляющую эмпирическую основу и обладают преимуществом предоставления интегрированного набора многомерных инструментов, предназначенных для работы с детьми и подростками.

Применение факторного анализа при разработке тестов

Пытаясь добиться систематической классификации черт личности, некоторые психологи обратились к факторному анализу. Этот метод, уже обсуждавшийся в связи с организацией когнитивных способностей, идеально подходит для сокращения числа категорий, необходимых для объяснения поведенческих феноменов путем поиска устойчивых паттернов в их появлении. Напомним, что элемент субъективности все же вносится в идентификацию факторов, поскольку этот процесс зависит от изучения мер или заданий (пунктов), имеющих наибольшую нагрузку по каждому фактору (см. главу 11). Отсюда, кросс-идентификация факторов из независимых исследований, в которых используются различные меры, становится трудноосуществимой, что всегда служило источником несоответствий как в названиях черт, так и в их количестве. Вдобавок ко всему, существует множество способов применения факторного анализа в исследовании черт личности. Среди этого множества есть две главные традиции, которые сосуществовали на протяжении нескольких десятилетий и между которыми в последние годы произошло некоторое сближение.

Одна из них, сосредоточенная на использовании данных личностных вопросников, иллюстрируется серией исследований Гилфорда и его сотрудников (см. Guilford, 1959, chap. 16; Guilford, & Zimmerman, 1956). Вместо того чтобы коррелировать суммарные показатели по существующим опросникам, эти исследователи вычисляли интеркорреляции между отдельными пунктами из многих личностных опросников. В качестве промежуточного продукта данного исследования были разработаны три личностных опросника, объединенные со временем в один инструмент — Гилфорда-Циммермана обследование темперамента (*Guilford-Zimmerman Temperament Survey*). Этот опросник дает отдельные показатели для 10 черт, а каждый показатель основывается на 30 разных пунктах. Примеры этих черт включают сдержанность (*Restraint*), доминирование (*Ascendance*), эмоциональную устойчивость и дружелюбность.

Еще одни новаторские приложения методов факторного анализа к разработке личностных опросников, иногда называемые «лексической» традицией, можно обнаружить в работе, начатой Р. Б. Кэттеллом в 1940-х гг. (John, Angleitner, & Ostendorf, 1988). Пытаясь добиться исчерпывающего описания личности, Кэттелл стал собирать все названия черт личности, встречающиеся или в специальном словаре (составленном Олпортом и Одбертом — Allport & Odbert, 1936), или в психиатрической и психологической литературе. Получившийся список из почти 18 000 терминов был сначала сокращен за счет объединения явных синонимов. Этот гораздо меньший перечень черт использовался затем при получении рейтингов знающих друг друга членов неоднородной группы взрослых людей. Вычисление интеркорреляций и факторный анализ этих рейтингов и данных стандартизованного самоотчета привел к установлению того, что Р. Кэттелл описал как «подлинно первичные черты личности» (*the primary source traits of personality*), — характеристика, которая, по-видимому, подразумевает большую универсальность и стабильность результатов, чем это представлялось оправданным на основании предыдущих исследований. Характерной особенностью подхода Р. Кэттелла является отношение к факторному анализу не как к способу сокращения размерности данных, а как к методу выявления базовых, причинных черт (Cattell, 1979).

Шестнадцатифакторный личностный опросник (16 PF). На основе факторных исследований Р. Кэттелл и его сотрудники разработали ряд личностных опросников, из которых наиболее известен Шестнадцатифакторный личностный опросник (*16 PF*), ныне доступный уже в своей пятой редакции (Cattell, Cattell, & Cattell, 1993; Conn, & Rieke, 1994; Russell, & Karol, 1994). Опубликованный впервые в 1949 г., *16 PF* предназначен для лиц в возрасте от 16 лет и старше и дает 16 показателей по таким чертам личности, как социальная смелость (*Social Bold*), доминантность (*Dominance*), вигильность (*Vigilance*), эмоциональная устойчивость (*Emotional Stability*) и власть сознания (*Rule Consciousness*). Эти 16 факторов, распознаваемые по тем же начальным буквам, уточнялись на протяжении ряда лет и, соответственно, переименовывались; со временем от той эзотерической терминологии, которую Кэттелл первоначально использовал для обозначения выделенных им черт личности, практически отказались. Например, края измерения (*dimension*), называемого теперь «Социальная смелость», сменили прежние, несколько загадочные ярлыки *Threctia* и *Parmia* на такие всем привычные маркеры, как «робость» и «смелость» соответственно. Факторы второго порядка в *16 PF*, число которых раньше варьировало от 4 до 8, сейчас называются «глобальными факторами» (*global factors*), а их количество ограничено 5 — числом, согласующимся с популярной пятифакторной моделью, которая будет рассмотрена в следующем разделе.

Пятая редакция *16 PF* доступна пользователям только в одной форме, содержащей 185 пунктов, большая часть которых была отобрана из предыдущих форм этого опросника на основе их содержания и добротности психометрических свойств. Для этой редакции *16 PF* установлены новые нормы на выборке 2500 лиц, отобранных таким образом, чтобы примерно отображать население США — согласно данным переписи 1990 г. — по параметрам пола, расы, возрастного состава и образования. Одной из отличительных особенностей *16 PF* является включение в него 15 пунктов, представленных в конце опросника в виде целостного блока заданий под заголовком «Вопросы на решение задач» (*Problem Solving Questions*); эти задания составляют Шкалу рассуждения (*Reasoning scale*), задуманную в качестве быстрой меры умственной способности. В добавление к этому, современная версия опросника дает три индекса стиля ответов, оценивающих уступчивость, выбор ответов наугад и попытки представить себя нереалистически, как обладающего социально желательными или, напротив, нежелательными качествами.

В пятой редакции опросника внутренняя согласованность и ретестовая надежность шкал для 16 первичных факторов выше, чем в его более ранних редакциях. Точно так же техническое руководство (*Technical Manual*) к пятой редакции *16 PF* содержит гораздо больше сведений о валидности, чем прежние руководства. Однако проблема отсутствия факторной независимости 16 первичных шкал, очевидная в ранних редакциях опросника, сохраняется, по всей видимости, и в его последней редакции. Эта проблема подчеркивается неспособностью всех без исключения исследователей, использовавших оригинальные переменные Кэттелла, воспроизвести его 16 факторов или получить хотя бы близкое факторное решение. Вместо этого в большинстве исследований, использующих данные, на которых Кэттелл построил свою систему, было выявлено от 4 до 7 факторов (Digman, 1990; L. R. Goldberg, 1993). Первые попытки воспроизвести первичные факторы Кэттелла были предприняты Фиске (D. W. Fiske, 1949) и привели к пятифакторному решению; эта работа сейчас широко цитируется как самый первый вариант обсуждаемой ниже современной модели.

«Пятифакторная модель» и почему она работает.[1] В современных работах по оценке личности все больше внимания уделяется так называемой «Пятифакторной модели» (*Five-Factor Model* или, сокращенно, *FFM*), которая символизирует необычайный уровень согласия среди исследователей личности, принадлежащих к самым разным школам факторно-аналитических исследований (Costa, & Widiger, 1994; Digman, 1990; McCrae, & John, 1992; Wiggins, & Pincus, 1992). В то же время конкретный вид, в котором эта модель представлялась, послужил причиной серьезной критики и полемики (Block, 1995; Carlson, 1992; Goldberg, 1993; Kroger, & Wood, 1993; Loevinger, 1994).[2] По существу, *FFM* — это попытка использовать иерархическую модель анализа с целью упростить ставшую необозримой коллекцию данных об эмоциональном поведении индивидуумов. Следовательно, она предоставляет информацию, с которой легче работать при оценивании конкретных людей и прогнозировании их поведения в определенных ситуациях. Факторы этой модели являются описательными, а не объяснительными, и они не относятся к разряду тех, что лежат в основе отдельных заданий или специфических тестов, из которых они извлечены.

Пятифакторная модель имеет сходство с иерархической структурой, получаемой в результате факторного анализа тестов способностей (см. главу 11). Хотя некоторые исследователи на протяжении последних четырех десятилетий сходились во мнении, что пять факторов — как раз то их число, которое необходимо для объяснения большинства корреляций между огромным количеством дескрипторов личности, наиболее подходящее для конкретных целей число факторов было бы лучше охарактеризовать как 5±2. К тому же и названия, даваемые каждому фактору, часто вызывают полемику (см., например, Digman, 1990, p. 423; Loevinger, 1994; Paunonen, 1993). Этого следовало ожидать, так как факторы отражают различия в выборе опросников, шкал, формата ответов, а также различия выборок, на которых они были получены.

Некоторая путаница и неправильное понимание были вызваны тем, как была описана процедура получения *FFM*. Сложилось впечатление, что *эти* базисные факторы личности были «открыты» благодаря некоему новому подходу. В действительности, эти факторы соответствуют второму уровню иерархии, получаемому в результате факторного анализа личностного теста и данных рейтинговой шкалы. Примечательно, что второй уровень иерархии, дающий факторы средней широты, оказался наиболее часто получаемым и воспроизводимым при оценивании как когнитивной, так и эмоциональной сферы. Поскольку они агрегируют или объединяют меры более низких уровней поведенческих дескрипторов и более узко определяемых черт, эти факторы дают более надежные показатели и, соответственно, более высокие коэффициенты валидности. Если тестовая батарея была сконструирована в соответствии с иерархической моделью и если есть нормы для разных уровней, как это имеет место в таких когнитивных батареях, как Дифференциальные шкалы способностей (см. главу 8), такую батарею можно применять максимально гибко, приспосабливая к конкретным целям. Например, после установления наиболее типичного для данного индивидуума

---

[1] Дополнительную информацию о «Пятифакторной модели» можно найти в двух недавно переведенных на русский язык учебниках: *Первин Л., Джон О.* Психология личности: Теория и исследования: Пер. с англ. — М.: Аспект Пресс, 2000 (гл. 8) и *Купер К.* Индивидуальные различия: Пер. с англ. — М.: Аспект Пресс, 2000 (с. 113–118). — *Примеч. науч. ред.*

[2] См. также другие статьи после заметки Kroger, & Wood в разделе «Критика» журнала *American Psychologist*, 1993, p. 1298–1304.

фактора среднего уровня, анализ можно дополнить изучением его показателей по более узким, более детальным тестам нижележащего уровня.

Два исследователя, чьи имена в первую очередь ассоциируются с Пятифакторной моделью, разработали тест, который согласуется с их версией этой модели. В своей современной редакции Пересмотренный *NEO*-личностный опросник[1] (*Revised NEO Personality Inventory* [*NEO PI-R*] — Costa, & McCrae, 1992b) дает показатели по пяти главным измерениям, или *областям (domains)*, личности и по 30 дополнительным чертам, или *аспектам (facets)*,[2] которые определяют границы каждой области. Пол Коста и Роберт МакКрей избегают употреблять термин «факторы» для обозначения шкал или компонентов, которые можно отнести к какому-либо уровню иерархии. Пять главных областей — Нейротизм (*Neuroticism* [*N*]), Экстраверсия (*Extraversion* [*E*]), Открытость опыту (*Openness to Experience* [*O*]), Уживчивость (*Agreeableness* [*A*]),[3] Сознательность (*Conscientiousness* [*C*]) — и их соответствующие аспекты приведены в табл. 13–1.

Шкалы *NEO PI-R* разрабатывались на протяжении 15-летнего исследования, которое началось с лонгитюдного изучения старения на выборках нормальных взрослых и позднее было распространено на клиническую, производственную и студенческую выборки. Хотя этот опросник создавался как средство измерения «черт нормальной личности», Коста и МакКрей предполагают, что он окажется полезным в клинической и прикладных областях деятельности, а также при проведении научных исследований. Среди методологических новшеств *NEO PI-R* — наличие формы для самоотчета (*Form S*) и двух вариантов формы для отчета стороннего наблюдателя (*Form R-Men* и *Form R-Women*), состоящей из тех же 240 пунктов, что и форма S, только сформулированных в третьем лице. Форма *R* дает возможность получить независимые оценки («рейтинги») от сверстников, супругов и других лиц по тем же областям, которые индивидуум оценивает сам у себя. Это особенно важно в случае *NEO PI-R*, поскольку этот опросник создавался в расчете на честного и сотрудничающего испытуемого и не содержит шкал, предназначенных для проверки правдивости ответов. Для взрослых мужчин и женщин нормы имеются по обеим формам; для лиц студенческого возраста обоих полов есть нормы только по форме *S*.

Пятифакторная модель (или «Большая пятерка») в том виде, как она постулирована Костой и МакКреем, получила широкое, хотя отнюдь не всеобщее, признание в качестве полезной теоретической рамки для изучения черт личности. Расхождения существуют даже среди сторонников факторно-аналитического подхода к исследова-

---

[1] Акроним *NEO* образован из первых букв слов *Neuroticism* (нейротизм), *Extraversion* (экстраверсия) и *Openness to Experience* (открытость опыту), однако сами эти слова не употребляются в полном названии опросника.

[2] Употребление терминов *domain* и *facet* (область и аспект, или домен и фацет соответственно) отражает все большее проникновение в психологию компьютерной терминологии, особенно связанной с разработкой искусственного интеллекта, и в частности экспертных систем, на основе сложных баз знаний. Каналом для такого проникновения, как нетрудно догадаться, служит когнитивная психология. — *Примеч. науч. ред.*

[3] Есть и другие варианты перевода этого термина: «доброжелательность» (М. С. Жамкочьян и В. С. Магун) и «готовность к согласию» (Т. М. Марютина и И. В. Равич-Щербо). Однако, как представляется, ни один из вариантов, включая и предлагаемый мною, не передает тех коннотаций этого английского слова, которые связаны с правилами поведения «истинного джентльмена». — *Примеч. науч. ред.*

Таблица 13–1

Области и аспекты Пересмотренного *NEO*-личностного опросника (*NEO PI-R*)

НЕЙРОТИЗМ	**ЭКСТРАВЕРСИЯ**
Тревожность (Anxiety [N 1])	Сердечность (Warmth [E 1])
Злобная враждебность (Angry Hostility [N 2])	Стадность (Gregariousness [E 2])
Депрессия (Depression [N 3])	Ассертивность (Assertiveness [E 3])
Застенчивость (Self-Consciousness [N 4])	Активность (Activity [E 4])
Импульсивность (Impulsiveness [N 5])	Поиск возбуждения (Excitement-Seeking [E 5])
Ранимость (Vulnerability [N 6])	Положительные эмоции (Positive Emotions [E 6])
ОТКРЫТОСТЬ ОПЫТУ	**УЖИВЧИВОСТЬ**
Воображение (Fantasy [O 1])	Доверие (Trust [A 1])
Эстетика (Aesthetics [O 2])	Прямота (Straitforwardness [A 2])
Чувства (Feelings [O 3])	Альтруизм (Altruism [A 3])
Действия (Actions [O 4])	Покладистость (Compliance [A 4])
Идеи (Ideas [O 5])	Скромность (Modesty [A 5])
Ценности (Values [O 6])	Мягкость (Tender-Mindedness [A 6])[1]
СОЗНАТЕЛЬНОСТЬ	
Компетентность (Competence [C 1])	
Порядок (Order [C 2])	
Чувство долга (Dutifulness [C 3])	
Достижение (Achievement [C 4])	
Самодисциплина (Self-Discipline [C 5])	
Осмотрительность (Deliberation [C 6])	

(С упрощениями из Costa & McCrae, 1992 b, p. 2. Copyright © 1992 by Psychological Assessment Resources, Inc. Воспроизводится с разрешения)

нию личности, — как в количестве предлагаемых ими факторов промежуточного уровня, так и в их определениях (L. R. Goldberg, 1993; Zuckerman, Kuhlman, Joireman, Teta, & Kraft, 1993). Тем не менее в своих различных вариантах, эта модель породила настоящий шквал исследовательской деятельности, нацеленной на кросс-идентификацию факторов и интеграцию таких разных перспектив, как значимые аспекты нормальной и патологической личности (см., например, Hofstee, de Raad, & Goldberg, 1992). Столь же быстрыми темпами продвигается работа по созданию тестов и усовершенствованию существующих шкал, имеющих отношение к «Большой пятерке» (Costa, & McCrae, 1994, 1995; Costa, & Widiger, 1994; Harkness, McNulty, & Ben-Porath, 1995; Hogan, & Hogan, 1992).

При оценивании результатов этой деятельности следует не забывать о том, что факторный анализ — это всего лишь способ объединения заданий (или пунктов) теста в относительно однородные и независимые кластеры. Такие группировки облегчают исследование валидности относительно эмпирических критериев, позволяют получить более эффективную комбинацию показателей для прогнозирования специфи-

[1] Этот полярный термин (антоним — *Tough-Mindedness*) У. Джеймса, впоследствии использованный Р. Б. Кэттеллом для характеристики полюсов одной из шкал *16 PF* (Harria—Premsia), имеет довольно сложное семантическое поле. См. его подробную характеристику в книге: *Джеймс У.* Воля к вере: Пер. с англ. — М.: Республика, 1997. — С. 209 и далее. — *Примеч. науч. ред.*

ческих критериев и способствуют более точному определению конструктов. Однородность и факторная чистота — желанные цели в конструировании тестов. Однако все это не заменяет (и не отменяет) эмпирическую валидизацию или исследование теоретических основ теста.

Теория личности в разработке тестов

Теории личности обычно зарождались в клинической среде. Объем экспериментальной проверки, которой они впоследствии подвергались, для разных теоретических систем крайне неодинаков. Независимо от степени такой объективной проверки, некоторое количество личностных тестов конструировалось в рамках той или иной теории личности. Сформулированные на клиническом материале гипотезы оказали особое влияние на разработку проективных методик, рассматриваемых в главе 15. Хотя при создании опросников, относящихся к классу стандартизованных самоотчетов, этому подходу следовали реже, некоторые примеры тому достаточно известны.

Клинический многоосевой опросник Миллона. Несмотря на то что Клинический многоосевой опросник Миллона-III (*Millon Clinical Multiaxial Inventory-III* [*MCMI-III*] — Millon, Millon, & Davis, 1994) в нескольких отношениях следует традиции *MMPI* и предназначен для тех же целей, этот инструмент содержит существенные методологические новшества. Фактически, его разработка была предпринята как обдуманная реакция на критику MMPI, опирающаяся на накопившиеся к этому времени достижения в диагностике психопатологии и конструировании тестов.

MCMI-III основывается на биопсихосоциальных (*biopsychosocial*) воззрениях Теодора Миллона в отношении нормального функционирования и психопатологии личности (Millon, 1969, 1981, 1990; Millon et al., 1996). Его теория очерчивает матрицу стилей личности (*personality styles*), выводимых путем комбинирования типов (*types*) по таким двум измерениям, как источник подкрепления (отдаленный, дискордантный, зависимый, независимый и амбивалентный) и характер совладающего поведения (активное или пассивное). Разработанная Миллоном теория стилей личности служит одним из концептуальных оснований новой формулировки категорий расстройств личности (Ось II) в Руководстве по диагностике и статистической классификации психических расстройств–III (*Diagnostic and Statistical Manual of Mental Disorders–III [DSM-III] — 1980*), подготовленном Американской психиатрической ассоциацией, теперь уже в четвертой редакции (*DSM-IV* — 1994). В свою очередь клинические шкалы *MCMI-III* совместимы, хотя и не полностью совпадают, с классификационной системой *DSM-IV*. Фактически, именно стремление поддерживать как можно более тесное соответствие между опросником и меняющейся со временем структурой *DSM* побуждало создателей *MCMI* к частым пересмотрам его шкал.

MCMI-III содержит 175 утверждений в форме самоописаний, которые респондент должен пометить либо как «верные», либо как «неверные». Профиль *MCMI-III* включает показатели по 24 клиническим шкалам, каждая из которых основана на частично перекрывающихся пунктах числом от 12 до 24, которые часто встречаются в различных шкалах, правда, не более чем в трех и притом с разными весами; пункты, отвечающие всем критериям валидизации для своей «родной» шкалы, получают весовой коэффициент, равный 2, тогда как вспомогательные пункты имеют весовой коэффици-

ент, равный 1. Как видно из табл. 13–2, клинические шкалы сгруппированы в четыре основные категории, а именно: Клинические паттерны личности (*Clinical Personality Patterns*), Тяжелая патология личности (*Severe Personality Pathology*), Клинические синдромы (*Clinical Syndromes*) и Тяжелые синдромы (*Severe Syndromes*). Первые две из этих категорий включают в себя шкалы, предназначенные для оценки устойчивых расстройств (Ось *II DSM*) паттернов личности. Две другие охватывают синдромы Оси *I DSM*. Есть еще три поправочных индекса и индекс валидности, назначение которых — обнаруживать атипичные паттерны ответов и систематические ошибки, обусловленные личными качествами тестируемых. Первоначально *MCMI-III* предполагал только компьютерную обработку результатов; в настоящее время, в дополнение к услугам обработки результатов по электронной почте и программному обеспечению для построения профилей и составления интерпретирующих отчетов, предоставляется документация для ручной обработки ответов на этот опросник, хотя она весьма трудоемка вследствие необходимости производить преобразование показателей и вводить ряд корректировок.

Одно из самых важных новшеств, внесенных *MCMI-III* в практику тестирования, — использование стандартных показателей, называемых показателями «базисного уровня» (*Base Rate* или, сокращенно, *BR*-показателями), которые, вместо нормализации, привязываются к преобладающим уровням измеряемых характеристик. Критические *BR*-показатели шкал *MCMI-III* устанавливаются таким образом, чтобы отражать актуарные данные для психиатрических популяций по базисным уровням специфических состояний, оцениваемых с помощью его шкал, и тем самым улучшать дифференциальную диагностику. Поскольку преобладающие уровни могут расходиться в разных клинических популяциях и условиях, *BR*-показатели некоторых шкал *MCMI-III* могут корректироваться в зависимости от условий жизни обследуемого, застарелости состояния, показателей по шкалам тревоги и депрессии, а также определенных паттернов ответов.

Разработка формулировок пунктов *MCMI-III* осуществлялась с позиций многоаспектного подхода, характерного для современной практики конструирования и валидизации личностных опросников. В этом отношении *MCMI-III* идет вразрез с методологией, описанной в нескольких разделах данной главы. Процедура включала три основных этапа: 1) независимый теоретический (формулирование и отбор пунктов опросника, соответствующих клинически релевантным конструктам), 2) внутренний структурный (вычисление корреляций «пункт — шкала» и подтверждение частот) и 3) внешний критериальный (например, установление отличий диагностических групп от контрольной группы и кросс-валидизация на новых выборках).

Контрольные (или эталонные) группы, используемые для анализа пунктов *MCMI-III* и его предшествующих версий, состояли из недифференцированных психиатрических больных, вместо того чтобы быть представленными выборками нормальных лиц. Главным образом вследствие исключительного использования клинических выборок при выведении норм и преобразованных показателей автор этого опросника ясно заявляет: «*MCMI-III* не является инструментом общего назначения, используемым для оценки личности в нормальных популяциях, и не предполагает никакого иного применения, кроме диагностического скрининга или клинического оценивания» (Millon et al., 1994, p. 5). К тому же критические показатели шкал и интерпретации профилей *MCMI-III* соотносятся с поведением тех, кто проявляет психопатологию средней степени тяжести, а не с реакциями тех, чьи проблемы либо близки проблемам нормаль-

Таблица 13–2

Шкалы Клинического многоосевого опросника Миллона-III

КЛИНИЧЕСКИЕ ПАТТЕРНЫ ЛИЧНОСТИ	КЛИНИЧЕСКИЕ СИНДРОМЫ
Шизоидный	Тревога
Избегающий	Соматоморфные расстройства
Депрессивный	Биполярный: маниакальный
Зависимый	Дистимия
Демонстративный	Алкогольной зависимости
Нарциссический	Наркотической зависимости
Антисоциальный	Посттравматического стрессового расстройства
Агрессивный (садистический)	
Компульсивный	ТЯЖЕЛЫЕ СИНДРОМЫ
Пассивно-агрессивный (негативистский)	Расстройства мышления
Обрекающий себя на неуспех (*Self-Defeating*)	Тяжелая депрессия (*Major Depression*)
	Бредовые расстройства

ТЯЖЕЛАЯ ПАТОЛОГИЯ ЛИЧНОСТИ	ПОПРАВОЧНЫЕ ИНДЕКСЫ	ИНДЕКС ВАЛИДНОСТИ
Шизотипическая	Раскрытия (*Disclosure*)	
Пограничная	Желательности (*Desirability*)	
Параноидная	Занижения (*Debasement*)	

(С упрощениями из Millon et al., 1994, p. 2. Copyright © by DICANDRIEN, INC. Воспроизводится с разрешения)

ных людей, либо относятся к крайне тяжелой патологии. В некоторых исследованиях *MCMI-III* с участием нормальных испытуемых обнаружилось, что их показатели оказались выше, чем у психиатрических больных, хотя и оставались на субклинических уровнях, тогда как в других исследованиях у нормальных людей по ряду шкал (например, демонстративности и нарциссизма) были получены показатели, относящиеся к патологическим уровням. Эти результаты подтверждают нецелесообразность применения этого опросника в работе с нормальными людьми и указывают даже на возможность того, что при превышениях среднего уровня эти шкалы измеряют качества здоровой личности (Wetzler, 1990).

Одно из основных назначений *MCMI-III* — служить вспомогательным средством в процессе дифференциальной диагностики. Эта задача осложняется фактом сосуществования у одного человека в одно и тоже время целого букета психиатрических состояний. Относительно частые пересмотры и усовершенствования, которым *MCMI* подвергался с момента его опубликования, достойны всяческого одобрения, но они же затрудняют оценку того, в какой степени этот опросник отвечает своему главному назначению. Однако большое количество исследований, проведенных с более ранними версиями *MCMI*, так же как и ряд связанных с ним публикаций, должны помочь его пользователям выйти из трудного положения. Результаты исследований свидетельствуют о том, что вхождение одних и тех же пунктов опросника в несколько шкал может уменьшать их различительную способность, особенно когда тестируемые оказываются в состояниях, относящихся к середине диапазона тревожности и депрессии. Не хватает и данных о диагностической эффективности шкал клинических синдро-

мов. Несмотря на это, *MCMI* обладает немалым потенциалом как инструмент для диагностики расстройств личности и оценки результатов их лечения (Choca, Shanley, & Van Denburg, 1992; Craig, 1993; Goncalves, Woodward, & Millon, 1994; Retzlaff, 1995; C. R. Reynolds, 1992a).[1]

Недавно Миллон разработал два новых инструмента, расширяющих его подход к оценке личности и психопатологии. Один из них — Подростковый клинический опросник Миллона (*Millon Adolescent Clinical Inventory* [*MACI*] — Millon, Millon, & Davis, 1993), которому отводится роль предпочтительного для использования инструмента при оценке подростков в возрасте от 13 до 19 лет в условиях клиники. MACI вырос из Подросткового личностного опросника Миллона (*Millon Adolescent Personality Inventory* [*MAPI*] — Millon, Green, & Meagher, 1982) — более старого инструмента, предназначенного как для проведения клинической оценки, так и для использования в профессиональном и образовательном консультировании, со шкалами, оценивающими основные стили личности, выраженные заботы и поведенческие тенденции подростков.[2] С другой стороны, Индекс стилей личности Миллона (*Millon Index of Personality Styles* [*MIPS*] — Millon, 1994) задуман как средство измерения личности взрослых, обращающихся к консультантам за помощью в решении рабочих, семейных и социальных проблем. *MIPS* стандартизован на выборках взрослых и студентов колледжей и сочетает в себе элементы теории личности Миллона с элементами теорий Фрейда и Юнга.

Список личных предпочтений Эдвардса. Среди теорий личности, стимулировавших развитие тестов, одной из наиболее результативных была система явных потребностей (*manifest needs*), предложенная Г. Мюрреем и его коллегами по Гарвардской психологической клинике (Murray et al., 1938). Одним из первых опросников, предназначенных оценить силу таких потребностей, был Список личных предпочтений Эдвардса (*Edwards Personal Preference Schedule* [*EPPS*] — Edwards, 1959). Начав с 15 потребностей, взятых из перечня Мюррея, Эдвардс подготовил набор утверждений, содержание которых, азалось соответствующим каждой из этих потребностей. Примеры включают потребности достижения (сделать все от тебя зависящее и выполнить нечто трудное), уважения (соответствовать тому, что от тебя ожидают), демонстрации (быть центром внимания), интрацепции (анализировать мотивы и чувства — свои собственные и окружающих), доминирования (влиять на других и считаться их лидером) и покровительства (помогать другим в затруднительных ситуациях).

Опросник состоит из 210 пар утверждений, в которых (парах) формулировки пунктов каждой из 15 шкал соотносятся с формулировками пунктов из остальных 14 шкал.[3] Из каждой пары тестируемые должны выбрать одно утверждение как более характерное для них. Важно иметь в виду, что благодаря этому вынужденному выбору *EPPS* дает *ипсативные (ipsative)* показатели, т. е. сила каждой потребности выражает-

---

[1] Карта обследования неадаптивной и адаптивной личности (*Schedule for Nonadaptive and Adaptive Personality* [*SNAP*]) — интересный новый измерительный инструмент, предназначенный для оценки личностной патологии. В противоположность *MCMI*, шкалы *SNAP* построены с помощью факторного анализа (см., например, Clark, McEwen, Collard, & Hickok, 1993).

[2] В настоящее время *MAPI* рекомендуется использовать только для неклинической оценки личности нормальных подростков.

[3] Эта форма задания, представляющая собой важную особенность *EPPS*, обсуждается в следующем разделе как пример методики вынужденного выбора.

ся не в абсолютных единицах, а относительно силы других потребностей индивидуума. Системой отсчета при ипсативном оценивании служит индивидуум, а не нормативная выборка. Так как сумма показателей по подшкалам постоянна для всех тестируемых, то если показатель по одной подшкале у конкретного испытуемого поднимается на одно деление вверх, его показатель по другой подшкале должен снизиться на одно деление. При этих условиях два испытуемых с одинаковыми показателями по *EPPS* могут заметно отличаться друг от друга по абсолютной силе своих потребностей. Несмотря на то что *EPPS* предоставляет нормы для перевода его показателей в процентили, целесообразность такого преобразования вызывает сомнения вследствие ипсативной природы показателей. Хотя ипсативная система отсчета, возможно, лучше всего подходит для *интраиндивидных (intraindividual)* сравнений, какие нужны, например, при оценке интересов и других предпочтений, данные, получаемые в нормативной системе отсчета, необходимы для такого вида *интериндивидных (interindividual)* сравнений, которые используются, к примеру, при оценке ненормальности (Fedorak, & Coles, 1979). Однако объединение этих двух систем отсчета в одном измерительном инструменте делает интерпретацию показателей путаной и менее выразительной по сравнению с тем, когда последовательно применяется какой-то один подход — ипсативный либо нормативный.

Хотя приводимые в руководстве к *EPPS* данные по валидности довольно скудны, опубликовано значительное количество независимых исследований по этому вопросу. И все же результаты этих исследований часто интерпретируются с трудом, поскольку в большинстве из них не учитывается ипсативная природа показателей *EPPS*. В случае ипсативных показателей средняя корреляция между отдельными шкалами имеет тенденцию быть отрицательной, а средняя корреляция всех шкал с любой внешней переменной стремится к нулю (Hicks, 1970). Вследствие этих искусственных ограничений ипсативные показатели невозможно должным образом проанализировать с помощью обычных корреляционных методов. Неудивительно поэтому, что опубликованные исследования по валидизации *EPPS* дали противоречивые и неинформативные результаты (см., например, Piedmont, McCrae, & Costa, 1992). Несмотря на свою простоту и ряд интересных конструктивных особенностей, *EPPS* нуждается в переработке для устранения технических недостатков, связанных с формой заданий и интерпретацией показателей.

Форма для исследования личности и другие опросники Джексона. Форма для исследования личности (*Personality Research Form* [*PRF*]) отражает многие технические достижения в конструировании тестов, в том числе некоторые процедуры отбора заданий (или пунктов опросников), которые были невозможны до появления быстродействующих вычислительных машин. *PRF* иллюстрирует подход Дугласа Н. Джексона к разработке личностных тестов, которая начинается с эксплицитного, подробного описания измеряемых конструктов. Эти описания составляют основу для формулирования пунктов опросника, а также для определения черт, которые предлагают для оценки экспертам в исследованиях валидности (Jackson, 1970, 1989b).

PRF доступна в пяти различных вариантах, включая два набора параллельных форм (*A*, *B* и *AA*, *BB*) из 300 и 440 пунктов соответственно. Более длинные формы дают показатели по 22 шкалам (по 20 пунктов каждая), в том числе два показателя валидности: Редкости (*Infrequency*) и Желательности (*Desirability*), тогда как более короткие формы имеют только 15 шкал (по 20 пунктов каждая). Разработанная позднее, с при-

менением более изощренных методов анализа заданий, дополнительная версия *PFR* (*Form E*) состоит из 352 лучших пунктов, отобранных из длинных форм опросника, и дает показатели по всем 22 шкалам (содержащих только по 16 пунктов каждая). Форма *E*, которая теперь используется чаще всех других, отличается от остальных вариантов *PRF* еще и более простой лексикой. Показатель Редкости, задуманный как индекс недобросовестности, непонимания инструкций и других не отвечающих цели опросника ответов, основывается на числе крайне невероятных ответов, выбираемых тестируемым, примеры которых включают такие формулировки, как: «Каждую ночь я пытаюсь хоть немного поспать», «Всю свою одежду и обувь я шью себе сам». Хотя систематическая ошибка желательности была существенно уменьшена благодаря использованию специальных методов еще на стадии разработки и отбора формулировок опросника, шкала Желательности все же сохранена в составе *PRF*. В руководстве к тесту корректно отмечено, что необычно высокие или низкие показатели по этой шкале могут указывать не только на атипичные аттитюды в отношении тестирования (например, преднамеренную попытку создать благоприятное или же, напротив, неблагоприятное впечатление), но также на важные характеристики личности (например, высокое самоуважение или высокую степень традиционной социализации в противоположность низкому самоуважению).

Подобно некоторым другим личностным методикам, PRF берет начало из теории личности Г. Мюррея. Основываясь на многочисленных исследованиях и обширной теоретической литературе последних трех десятилетий, Д. Джексон сформулировал поведенчески ориентированные и взаимоисключающие определения 20 конструктов, или черт личности. Из них 12 имеют те же названия, что и конструкты, положенные в основу *EPPS*. Для каждой из 20 биполярных личностных шкал в руководстве к тесту приводится описание лиц с высокими показателями по данной шкале и набор определяющих соответствующую черту характеристик. Два примера определений шкал приведены в табл. 13–3.

С помощью тщательно контролируемых процедур по каждой шкале было создано более 100 пунктов-утверждений. Затем на основе высокой бисериальной корреляции с суммарным показателем шкалы и низкой корреляции с показателями остальных шкал черт личности и шкалы Желательности было отобрано по 20 пунктов для каждой шкалы опросника. Кроме того, были исключены пункты с крайними частотами ответов. С помощью специально разработанной компьютерной программы пункты распределялись по двум параллельным формам опросника исходя из их бисериальной корреляции со своей шкалой и частоты ответов. При конструировании формы *E* в число используемых методов входило вычисление индекса эффективности задания (*item efficiency index*) для каждого пункта опросника. Основанный на весовых коэффициентах, выводимых из разнообразных статистических параметров задания, этот индекс позволяет определить ранг каждого пункта внутри своей шкалы исходя из его эффективности.

Конструктная валидность *PRF* зависит в значительной степени от процедур, используемых при разработке и отборе заданий для каждой шкалы. Проведенный впоследствии факторный анализ подтвердил правомерность группировки заданий по 20 шкалам. Корреляции с сопоставимыми шкалами Калифорнийского психологического опросника, Гилфорда—Циммермана обследования темперамента и *NEO-PI*, помимо прочих, служат дополнительным подтверждением того, что *PRF* выявляет эти черты. Хотя исследования корреляций между различными инструментами для изме-

Таблица 13–3

Примеры определений шкал из Формы для исследования личности (*PRF*)

Шкала	Описание тестируемого с высоким показателем	Характеристики измеряемой черты
Склад познания	Не любит двусмысленности или неопределенности в информации; хочет на все вопросы получать полные ответы; стремится принимать решения, основываясь на точном знании, а не на догадках или вероятностных предположениях.	Точный, требовательный, определенный, стремящийся к достоверности, дотошный, добивающийся во всем совершенства, разъясняющий все до мелочей, прямой, аккуратный, доскональный, буквальный, избегающий двусмысленности, устанавливающий четкие границы, негибкий, нуждающийся в структуре.
Склад чувствительности	Замечает оттенки запахов, звуков, зрительных и вкусовых ощущений, обращает внимание на то, как воспринимаются вещи; помнит свои ощущения и считает их важной частью своей жизни; чувствителен ко многим видам чувственного опыта; может придерживаться по сути гедонистического или эстетического взгляда на жизнь	Эстет, получает удовольствие от телесных ощущений, наблюдательный, реалистичный, сознающий происходящее, обращающий внимание на окружающую обстановку, имеющий вкус, чувствительный, чувственный, открытый опыту, восприимчивый, замечающий, тонко различающий, полон впечатлений.

(Из Jackson, 1989 b, pp. 6–7. Copyright © by Douglas N. Jackson. Воспроизводится с разрешения)

рения потребностей по Мюррею — *PRF, EPPS, TAT* (глава 15) и *ACL* (глава 16) — дали противоречивые результаты, полученные в них данные все же в большей мере подтверждают конструктную валидность *PRF*, чем некоторых других инструментов. И это неудивительно, если не забывать об особом внимании, которое уделялось формулированию черт при разработке *PRF* (Costa, & McCrae, 1988; D. W. Fiske, 1973; Rezmovic, & Rezmovic, 1980). Сведения по эмпирической валидности PRF относительно самооценок и оценок, даваемых другими людьми, также выглядят многообещающими. Заслуживает упоминания и то обстоятельство, что *PRF* оказался пригодным для использования в нескольких культурах, включая культуры незападного типа (Jackson, Guthrie, Astilla, & Elwood, 1983; Paunonen, Jackson, Trzebinski, & Forsterling, 1992). В целом, как свидетельствует составленная МакЛеннаном (McLennan, 1992) обширная аннотированная библиография, *PRF* является отличным исследовательским инструментом, однако для определения его эффективности в практических ситуациях требуется дополнительная информация.[1]

Пересмотренный личностный опросник Джексона, разработанный после *PRF* с помощью аналогичных, хотя и более тонких методов конструирования шкал, имеет более практическую ориентацию (Jackson, 1976, 1994a). Шкалы черт отбирались из ли-

[1] Обзоры, посвященные применению этого инструмента, часто публиковались в Ежегодниках психических измерений: 7th MMY, # 123; 8th MMY, # 643; 10th MMY, # 282; см. также TIP-IV, в котором помещен обновленный список ссылок.

тературы по психологии личности и социальной психологии отчасти в силу их отношения к предсказанию поведения нормальных людей в разнообразных контекстах. Среди черт, охватываемых 15 шкалами этого опросника, — тревожность, склонность к сотрудничеству, ответственность, социальная прозорливость (*social astuteness*) и толерантность. Данные о валидности собирались не только путем установления корреляций с самооценками и оценками других людей на основе матричной модели «черты — методы», но и посредством изучения специальных групп, в отношении которых имелись релевантные данные о поведении в реальных жизненных ситуациях. При последнем пересмотре этого опросника были обновлены нормы для студентов колледжей и разработаны новые нормы для «белых и синих воротничков», с тем чтобы обеспечить прочную основу для его применения в учебном консультировании и кадровой работе. В добавление к этому была произведена техническая доработка шкал с внесением в них незначительных изменений, и теперь все шкалы сгруппированы в пять кластеров более высокого порядка, которые, в основном, совместимы с категориями обсуждавшейся выше Пятифакторной модели (*FFM*). В частности, такие кластеры *JPI-R*, как «Экстравертированный» (*Extroverted*), «Надежный, заслуживающий доверия» (*Dependable*) и «Аналитический» (*Analytical*), имеют существенное сходство с такими измерениями *FFM*, как «Экстраверсия», «Сознательность» и «Открытость опыту» соответственно. Вдобавок, кластер «Эмоциональный» (*Emotional*) в *JPI-R*, по-видимому, представляет собой комбинацию таких измерений *FFM*, как «Нейротизм» и «Уживчивость». В то же время кластер *JPI-R* «Оппортунистический» (*Opportunistic*), полученный из шкал «Социальной прозорливости» и «Склонности к риску», не имеет прямой связи с *FFM* и потому может считаться характерным фактором (*unique factor*).

Обращаясь к оценке психопатологии, Джексон применил те же строгие стандарты, что использовались при создании *PRF* и *JPI*, к конструированию Базисного личностного опросника (*Basic Personality Inventory* [*BPI*] — Jackson, 1989a). *BPI*, разработка которого заняла примерно 15 лет, имеет целью воссоздать диагностическую эффективность, связываемую с *MMPI*, за счет использования более основательных — с точки зрения содержания, психометрической чистоты и диапазона применимости — шкал. Хотя *BPI* требует установления более репрезентативных норм, особенно для взрослых, он уже хорошо зарекомендовал себя в качестве инструмента для клинической оценки в области подростковой делинквентности (Holden, & Jackson, 1992; что касается критического обзора, см. Urbina, 1995).

В этом обзоре опросников, относящихся к классу стандартизованных самоотчетов, читатель, возможно, обратил внимание на растущую тенденцию сочетать при их разработке различные подходы. Это особенно справедливо в отношении опросников Джексона, да и других новых опросников, созданных в последнее десятилетие, при разработке которых используются все стратегии, за исключением привязки пунктов опросника к эмпирическим критериям. Даже семейство Миннесотских опросников, служащих выдающимися примерами применения процедуры привязки к эмпирическому критерию, в настоящее время содержат шкалы, основанные на результатах контент-анализа и факторного анализа. Некоторые данные свидетельствуют о том, что шкалы личностных опросников, сконструированные с помощью любого из четырех основных методов, описанных в этой главе, могут быть эффективными, по крайней мере в том, что касается конвергентной и прогностической валидности (Burisch, 1986). Однако опирающиеся на содержание и на теорию процедуры легче в построении, эффективнее в использовании и с большей вероятностью показывают дискриминантную валидность,

чем метод привязки к эмпирическому критерию, который, в свою очередь, по крайней мере сравним с оставшимся. К тому же большинство сходится в том, что: а) разработку опросников следует начинать с эксплицитного определения подлежащих измерению черты или конструкта и что б) организационная структура, намеченная Кэмпбеллом и Фиске в их матрице «свойства — методы», обеспечивает оптимальную стратегию для проверки конструктной валидности личностных опросников (Angleitner, John, & Löhr, 1986; Hogan, & Nicholson, 1988; Ozer, & Reise, 1994).

Аттитюды тестируемых и систематическая ошибка в ответах

Фальсификация и социальная желательность. Опросники типа стандартизованных самоотчетов особенно подвержены возможности искажения или фальсификации сообщаемых о себе сведений. Несмотря на вступительные замечания об отсутствии «единственно правильных ответов», большинство пунктов в таких опросниках предполагают один ответ, признаваемый как социально более желательный или приемлемый, чем другие. В таких тестах респонденты, если они устраиваются на работу или поступают на учебу, могут испытывать искушение «прикинуться хорошими» или выбирать ответы, которые создают благоприятное впечатление. При других обстоятельствах респондентами может руководить стремление «притвориться плохими», показывая себя психологически более неблагополучными, чем это есть на самом деле. Такое может произойти, например, при тестировании лиц, привлеченных к суду за уголовное преступление.

Доказательств тому, с каким успехом респонденты могут притворяться при ответах на личностные опросники, более чем достаточно (см., например, Jacobs, & Barron, 1968; Radcliffe, 1966; Stricker, 1969; J. S. Wiggins, 1966). Этот факт можно легко продемонстрировать в любой студенческой аудитории, попросив различные группы принять определенные роли. Например, часть аудитории попросить руководствоваться в своих ответах на вопросы тем, как ответил бы на них удачливый и уравновешенный студент колледжа; другую часть — тем, как ответил бы на них плохо адаптированный человек; остальных попросить отвечать на вопросы о собственном поведении искренне. Можно также одним и тем же людям предъявить тест дважды: в первом случае с инструкцией симулировать в ответах какой-то определенный тип поведения, а во втором — отвечать правдиво. Результаты этих исследований есть яркое свидетельство той легкости, с которой умышленно создается при ответах на такие опросники желаемое впечатление. Интересно отметить, что не менее успешно может симулироваться профпригодность в ситуациях отбора кандидатов для конкретной работы (Wesman, 1952). Современные исследования показывают, что, особенно при использовании опросников типа стандартизованных самоотчетов, «очевидная валидность» (см. главу 5) задания увеличивает его подверженность фальсификации, причем как в реальных, так и в контролируемых лабораторных условиях. Чем легче респондентам распознать черту, оцениваемую по словесной формулировке пункта опросника, тем чаще они дают желательный ответ (Bornstein et al., 1994).

Тенденция к выбору социально желательных ответов в опросниках типа стандартизованных самоотчетов необязательно означает сознательный обман со стороны от-

вечавшего. Эдвардс (A. L. Edwards, 1957), первым исследовавший такую переменную, как социальная желательность, концептуализировал ее преимущественно как эффект фасада или тенденцию «выдвигать на первый план хорошие стороны», которую респонденты по большей части не сознавали. Эта тенденция может указывать на недостаточное проникновение в свои личные особенности, самообман или неготовность примириться с собственными недостатками. Другие исследователи (Crowne, & Marlowe, 1964; N. Frederiksen, 1965) представили доказательства в пользу того, что сила установки на социально желательные ответы связана с более общей потребностью индивидуума в самозащите, желанием избежать критики, социальной конформностью и стремлением заслужить социальное одобрение. С другой стороны, человеком, выбирающим неблагоприятные для описания самого себя утверждения, может двигать потребность во внимании, сочувствии или помощи в решении личных проблем. Например, заинтересованные в психотерапии лица при заполнении личностного опросника, по-видимому, могут показать себя менее приспособленными к действительности, чем это есть на самом деле.

К тому же не следует думать, что фундаментальные исследования свободны от влияния установок на ответы. В контексте изучения изменения аттитюдов, например, исследователи показали, что на результаты могут влиять такие условия, как чувствительность испытуемого к ожиданиям экспериментатора, желание сохранить свое лицо, стремление угодить экспериментатору или сорвать его планы (Silverman, & Shulman, 1970). Непредвиденными расхождениями в этих установках на ответы можно отчасти объяснить неполную воспроизводимость результатов при повторении экспериментов.

Некоторые исследователи (Paulhus, 1984, 1986; Paulhus, & Reid, 1991) подчеркивали различие между понятиями «создания впечатления путем обмана» (*impression management*)[1] и «самообмана» (*self-deception*) как толкованиями социально желательного — или нежелательного — реагирования. ***Создание впечатления путем обмана*** указывает на сознательное введение в заблуждение других с целью вызвать специфический эффект, желательный для респондента. Этот вид реагирования рассматривается как искажение данных самоотчета и как нечто такое, что само по себе должно быть оценено и, по возможности, сведено к минимуму или ограничено. С другой стороны, самообман обычно состоит в смещении ответов тестируемого, которые сам он считает совершенно верными, в положительную сторону и представляет собой гораздо более сложный феномен. Самообман связан с другими понятиями, имеющими отношение к «Я», такими как Я-образ (*self-image*) и самоуважение (*self-esteem*), а также с психоаналитическим понятием защитных механизмов. Поэтому эта переменная заслуживает изучения сама по себе, — как указывающая на хорошее приспособление к действительности, разумеется до известной степени, и как предсказывающая другие независимые критерии. Отмечалось, например, что некоторые шкалы стандартизованных самоотчетов могут давать результаты, которые показывают «иллюзорное психическое здоровье» у тех, кто с помощью такой защиты, как отрицание (*defensive deniers*) стремится сохранить веру в свою приспособленность к жизни (см., например, Shedler, Mayman, & Manis, 1993). Таким образом, связь между самообманом и приспособлением к действительности, скорее всего, не является простой, как, впрочем, и прямой. Вопрос еще больше усложняется тем обстоятельством, что, по-видимому, взаимодей-

[1] Вместо *impression management* иногда используются такие термины, как *manipulation* (манипулирование, подтасовка) и *other-deception* (обман других, жульничество).

ствие ряда лингвистических характеристик формулировок пунктов опросника с переменными респондента может вызывать систематическую ошибку в ответах (Helfrich, 1986).

Для решения проблемы фальсификации и связанной с ней проблемы установок на ответ применялось несколько методик. В некоторых опросниках действие этих факторов можно снизить за счет конструирования относительно «тонких» или социально нейтральных формулировок их пунктов, однако, как показал Джексон (Jackson, 1971), часто именно эти пункты опросника имеют низкую валидность относительно измеряемой переменной. В идеале, да и в некоторых реальных ситуациях, тестовые инструкции и установление раппорта должны мотивировать тестируемых отвечать искренне, если они убеждены в том, что такие ответы пойдут им на пользу. Однако в определенных ситуациях такой подход оказывается неэффективным и, вероятно, не оказывает сколько-нибудь существенного влияния на несознаваемые тестируемыми установки отвечать в социально желательном ключе.

Другой подход к оценке социально желательного реагирования и других форм создания впечатления путем обмана состоит в конструировании специальных шкал, которые могут либо встраиваться в опросник, либо предъявляться в виде самостоятельных шкал в составе батареи тестов. Одной из самых первых шкал такого типа была шкала социальной желательности (*SD*) Эдвардса (Edwards, 1957), разработанная путем отбора пунктов опросника на основе сходства мнений экспертов относительно крайне высокой или низкой социальной желательности их формулировок. В других шкалах, таких как Шкала благоприятного впечатления из *CPI* и шкала *SD* Виггинса (J. S. Wiggins), формулировки отбирались на основе различий в частоте одобряемых ответов у тех, кто заполнял опросник в условиях, требующих «прикинуться хорошим», и у тех, кто заполнял опросник в нормальных условиях. Иллюстрацией третьего метода может служить входящая в состав *MMPI* шкала лжи, которая включает в себя пункты, сформулированные таким образом, что отвечать на них в социально желательном ключе будут лишь те респонденты, которые стремятся выставить себя нереалистически положительными. Придумывались и другие меры, нацеленные специально на обнаружение симуляции, недобросовестности или случайного характера ответов при заполнении опросников, такие, например, как F-шкала в *MMPI*.[1] Еще один метод, нацеленный, однако, не на обнаружение, а на предотвращение симуляции, состоит в использовании заданий с вынужденным выбором.

Методика вынужденного выбора. В сущности, эта методика требует от респондента сделать выбор между двумя описательными терминами или фразами, которые выглядят одинаково приемлемыми, но различаются валидностью. Такие спаренные фразы в социальном отношении могут быть либо обе желательными, либо обе нежелательными. Задания, построенные по типу вынужденного выбора, могут состоять из 3, 4 или даже 5 пунктов. В таких случаях респонденты должны указать, какая из фраз является для них наиболее, а какая наименее характерной. В еще одном варианте требуется выбрать между двумя контрастирующими ответами в рамках одной и той же черты, оцениваемыми по единой шкале. Хотя эта форма заданий редко используется в личностных опросниках, ее преимущество в том, что она обеспечивает получение нор-

[1] Исчерпывающий обзор шкал для оценки симуляции при заполнении опросников дан в работе Berry, Wetter, & Baer (1995).

мативных, а не ипсативных показателей, и потому не накладывает искусственных ограничений на взаимосвязи между разными шкалами. В качестве примера можно назвать Индикатор типов Майерс—Бриггс, обсуждаемый в главе 16.

Использование методики вынужденного выбора для контроля социальной желательности требует двух типов сведений о каждой альтернативе ответа, а именно информации о ее социальной желательности (или знания «индекса предпочтения») и информации о ее валидности (или знания «индекса различения»). Последний может определяться на основе любого конкретного критерия, для прогнозирования которого предназначен опросник, такого как академическая успеваемость или успех в конкретной работе; или же этот индекс может основываться на факторных нагрузках выбираемых ответов либо на их теоретической релевантности различным чертам личности. Социальную желательность можно определить на основе оценок репрезентативной группой альтернатив ответов в отношении данной переменной или путем подсчета частоты, с которой эти альтернативы встречаются в самоописаниях. Было показано, что частота выбора и оцененная социальная желательность очень сильно коррелируют (Edwards, 1957). Другими словами, *усредненное* самоописание популяции тесно согласуется с ее усредненным описанием желательной личности.

Хотя влияние фактора социальной желательности можно уменьшить благодаря использованию заданий с вынужденным выбором, вряд ли можно рассчитывать на его полное исключение. Когда *EPPS* предъявлялся в формате свободного выбора, его показатели довольно высоко коррелировали с показателями, полученными при предъявлении того же теста в формате вынужденного выбора (Lanyon, 1966). Более того, оцененная социальная желательность конкретных формулировок пунктов опросника не является раз и навсегда заданной величиной, но может различаться для лиц, преследующих разные професиональные или учебные цели. Поэтому тесты, построенные по типу вынужденного выбора, задания которых уравнивались относительно общей социальной желательности, все равно могли фальсифицироваться, когда на них отвечали претенденты на получение работы или абитуриенты университетских профессиональных школ, а также другие группы людей, преследующих конкретные цели. С другой стороны, было установлено, что когда формулировки пунктов опросника объединяют в пары на основе усредненных *групповых* суждений об общей социальной желательности, они могут оказаться далеко не одинаковыми для *индивидуумов* (N. Wiggins, 1966).

Итак, в том, что касается контроля фальсификации или установок на социальную желательность ответа, методика вынужденного выбора, по-видимому, оказалась не столь эффективной, как ожидалось. В то же время формат заданий с вынужденным выбором, особенно когда он используется для получения ипсативных показателей, вносит дополнительные технические трудности и не позволяет получить информацию об абсолютной силе индивидуальных особенностей, которая может иметь первостепенное значение в некоторых ситуациях тестирования.

Установки на ответ и стили ответов. Тенденция выбрать альтернативы ответа на основе социальной желательности — это только одна из нескольких установок на ответ, которые были выявлены при изучении ответов на опросники типа стандартизованных самоотчетов (Lanyon, & Goodstein, 1982, p. 158–169). Хотя многочисленная литература по действию установок на ответ в личностных опросниках появилась в основном после 1950-х гг., о влиянии установок на ответ в тестах способностей и личности писалось и раньше (см. Block, 1965, chap. 2). Одной из установок на ответ, давно

привлекшей к себе внимание исследователей, была установка на *согласие (acquiescence)*, или тенденция отвечать «верно» или «да». Такое молчаливое согласие теоретически мыслится как непрерывная переменная: на одном конце шкалы находится последовательный «соглашатель», а на другом — последовательный «отрицатель» (Couch, & Keniston, 1960). При конструировании личностных опросников эта установка на ответ учитывается тем, что число пунктов шкалы, на которые ответ «да» или «верно» засчитывается как признак наличия измеряемой с ее помощью черты, должно быть равно числу пунктов, для которых таким ответом будет «нет» или «неверно». Такой сбалансированности можно достигнуть правильным отбором и переформулированием пунктов шкал, как это было сделано, например, в *PFR* и в настоящее время делается в новейших опросниках.[1]

Еще одной достаточно исследованной установкой на ответ является *отклонение (deviation)*, или тенденция давать необычные или необщепринятые ответы. Берг (Berg, 1967) выдвинул эту гипотезу и продемонстрировал действие фактора отклонения на невербальном содержании специально разработанного теста, в котором требовалось выразить предпочтение одной из геометрических фигур. Шкалы, составленные из пунктов, на которые почти все тестируемые будут, вероятно, давать ответы одного типа, такие как шкала Редкости в *PRF* Джексона, как раз и предназначены для выявления отклоняющихся паттернов ответов. Однако сам Джексон, помимо других, указывал на то, что эти шкалы не всегда концептуально связаны с внешним критерием и потому порождают новые проблемы, особенно в таких ситуациях, как оценка профпригодности, где релевантность пунктов опросника содержанию профессиональной деятельности считается весьма важным условием. Именно по этой причине при пересмотре такого опросника, как *JPI*, из него была исключена шкала Редкости (Jackson, 1994a). Тенденция использовать крайние точки на рейтинговой шкале (например, 1-ю и 7-ю на шкале с семью делениями) также была зафиксирована как возможная систематическая ошибка в ответах (Paulhus, 1991).

Исследование таких установок на ответ, как социальная желательность, согласие и отклонение, прошло несколько этапов. Кода установки на ответ были впервые обнаружены, они рассматривались как источник нерелевантной дисперсии или, иначе говоря, дисперсии ошибки, подлежащей устранению из тестовых показателей. Позднее эти установки на ответ стали рассматриваться как индикаторы широких и прочных характеристик личности, которые сами по себе заслуживали исследований и измерений (Jackson, & Messick, 1958, 1962; J. S. Wiggins, 1962). На этом этапе они обычно характеризовались как *стили ответов (response styles)*, и в отношении них скрупулезно накапливались эмпирические данные. В конченом итоге эти данные были подвергнуты критике с разных точек зрения (Block, 1965; Heilbrun, 1964; Rorer, 1965). Блок (Block, 1965), например, представил убедительное доказательство правомерности содержательно-ориентированной интерпретации двух главных факторов, обычно выявляемых при объяснении большей части общей дисперсии в шкалах *MMPI*, которые (факторы) сторонники введения понятий «установка на ответ» и «стиль ответа» интерпретировали как социальную желательность и согласие.

[1] Хелмс и Реддон (Helmes, & Reddon, 1993) также показали, что в случае биполярных шкал с несбалансированными по ключу (да/нет) пунктами объем информации, передаваемый низкими показателями, существенно снижается.

Спорные вопросы установок на ответ и ставшая классической контроверза «содержание/стиль» в оценке личности так и не были полностью разрешены (Edwards, 1990; Hogan, & Nicholson, 1988; Jackson, & Paunonen, 1980).[1] Большинство разработчиков тестов и исследователей, по-видимому, сходятся в том, что показатели личностных опросников скорее всего отражают своеобразную комбинацию самообмана, попыток обмануть других и реалистического изображения себя и что вес каждого из этих компонентов варьируется в зависимости от того, кто конкретно и при каких обстоятельствах тестируется. Некоторые, однако, расценивают попытки повысить достоверность данных самоотчета путем создания специальных шкал и заданий как приводящие к обратным результатам в том смысле, что это может снижать валидность шкал, особенно в отношении нормальных выборок в противоположность патологическим. Такие авторы пропагандируют использование навыков клинической работы, чтобы склонить пациента к сотрудничеству и адекватно проинтерпретировать показатели, а также дополнение тестовых данных рейтинговыми оценками от осведомленных информантов всякий раз, когда имеются основания подозревать серьезные искажения (см., например, Costa, & McCrae, 1992a).

Большинство других работников, особенно связанных с оцениванием психопатологии, продолжают использовать так называемые шкалы «валидности», сознавая при этом, что они могут также отражать стили и характеристики личности. Фактически, в некоторых из новейших и технически более совершенных инструментов для оценки психопатологии, таких, например, как *BPI* Джексона и *PAI* Мори, используются *и* сбалансированные по ключу формулировки пунктов, *и* специальные шкалы для обнаружения установок на ответы, снижающие валидность. Уже образовалось новое множество таких шкал, примерами которых могут служить шкалы *VRIN* и *TRIN* из *MMPI-2* и *MMPI-A*, построенных из специально подобранных пар сходных либо противоположных по содержанию утверждений и используемых для обнаружения непоследовательных или противоречивых ответов. Благодаря такому способу их построения шкалы *VRIN* и *TRIN*, которые сходны с предложенной Грином (Greene, 1978) шкалой Недобросовестности (*Carelessness scale*) для оригинального *MMPI*, вероятно, свободны от смешанной дисперсии валидных черт личности (Ozer, & Reise, 1994).

В любом случае, споры вокруг установок на ответ и стилей реагирования стимулировали широкие исследования и вызвали к жизни несколько сотен публикаций. Подобно многим научным спорам, эта длительная полемика привела к заострению нашего внимания на методологических проблемах, что позволило лучше разобраться в них и, благодаря этому, усовершенствовать конструирование личностных опросников и их использование для решения научно-исследовательских и практических задач.

Черты, состояния, люди и ситуации

Взаимодействие между человеком и ситуацией. Давнишний спор между сторонниками универсальности психологических черт и защитниками ситуационной специфичности поведения достиг своего пика в конце 1960-х — начале 1970-х гг. Ряд событий 1960-х гг. переместил фокус внимания с широко определяемых черт к узко опре-

[1] См. также другие статьи, следующие за статьей Эдвардса в том же разделе «Критика» журнала *American Psychologist*, 1990, pp. 289–295.

деляемому «заинтересованному поведению» (*behaviors of interests*). В сфере способностей этот новый фокус внимания иллюстрируется индивидуализированными учебными программами и предметно-ориентированным тестированием (глава 3), так же как диагностикой и программами преодоления трудностей в обучении (глава 17). В сфере личности сильнейший толчок к изучению специфичности поведения дали теории социального научения и социально-когнитивные теории, лежащие в основе модификации поведения и поведенческой терапии (Bandura, 1969, 1986; Goldfried, & Kent, 1972; Michel, 1968, 1969, 1973). Критика была нацелена преимущественно на ранние представления о чертах как постоянных, не поддающихся изменению внутренних причинных сущностях. Такого рода критика в отношении всех черт, — как когнитивных, так и некогнитивных, — уже предугадывалась в более ранних исследованиях и публикациях некоторых психологов (см. главу 11). Хотя лишь немногие отстаивали эту крайнюю точку зрения на черты, в самый разгар споров о ситуационной специфичности поведения трудно было найти такого человека, который бы открыто признал себя «твердо стоящим на позиции теории черт» (Jackson, & Paunonen, 1980).

Ситуационная специфичность, в частности, гораздо более характерна для черт личности, чем для способностей. Так, человек может быть достаточно общительным и открытым на работе, но застенчивым и замкнутым на дружеской вечеринке, или же студент, пользующийся шпаргалками на экзаменах, может быть абсолютно честным в денежных делах. Обширные эмпирические данные, собранные Мишелем (Michel, 1968) и Д. Петерсоном (D. Peterson, 1968), показывают, что люди действительно проявляют подобную ситуационную специфичность во многих неинтеллектуальных аспектах поведения, таких как агрессивность, социальная конформность, зависимость, твердость, честность и аттитюды в отношении власти. Частичное объяснение более высокой межситуационной устойчивости когнитивных функций по сравнению с некогнитивными можно отыскать в большей стандартизации реакций индивидуума в интеллектуальной области по сравнению с личностной сферой (Anastasi, 1958, chap. 11; 1970, 1983a). Например, обязательная школьная программа вносит существенный вклад в развитие широко применяемых когнитивных навыков в вербальной и числовой областях. Развитие же личности происходит в гораздо менее единообразных условиях. Более того, в области личности одна и та же реакция может вызвать социальные последствия, положительно подкрепляемые в одной ситуации и отрицательно — в другой. Это означает, что индивидуум может научиться реагировать по-разному в разных контекстах. Кроме того, такие различия жизненного опыта, связанные с различиями конкретных людей, ситуаций и культур, приводят к большей неоднозначности формулировок пунктов личностных тестов по сравнению с заданиями когнитивных тестов. Поэтому один и тот же ответ на конкретный вопрос личностного опросника, который сам по себе может истолковываться как некая «ситуация», возможно, будет иметь различное диагностическое значение для разных испытуемых.

Следует отметить, что контроверза «черта/ситуация» связана с уже знакомой нам проблемой соотношения наследственности и среды (D. C. Rowe, 1987). Факторы наследственности, вероятно, проявляются в относительно устойчивых индивидуальных чертах, которые тем не менее могут включать и такую черту, как приспособляемость к ситуационным требованиям. Средовые факторы могут вносить вклад как в ситуационную изменчивость (специфичность), так и в устойчивость черты, поскольку сама среда индивидуума может демонстрировать значительное временное и ситуационное постоянство. Соответствующие экспериментальные планы с повторными

измерениями, проводимыми через определенные промежутки времени (продольно) и в различных ситуациях (поперечно), должны улучшить наше понимание обеих проблем.

Теоретические дискуссии и эмпирическое изучение взаимодействия «человек/ситуация» безусловно обогатили наше понимание многих условий, которые определяют индивидуальное поведение, и внесли существенный вклад в разработку изощренных планов исследований. Одновременно отмечено растущее согласие между приверженцами противоположных взглядов в отношении того, что объяснения поведения чертами личности могут сосуществовать с его ситуационными объяснениями и что в действительности индивидуальное поведение детерминировано взаимодействием черт и ситуационных переменных. Сближение позиций произошло прежде всего в ходе спокойного и содержательного обсуждения данной проблемы в различных изданиях в период с конца 1970-х до конца 1980-х гг.[1] В этих дискуссиях выяснилось несколько заслуживающих упоминания вопросов. Поведение обнаруживает значительную временную устойчивость в тех случаях, когда оно надежно измеряется, а именно путем суммирования повторных наблюдений предпочтительно многих наблюдателей, хорошо знающих оцениваемого человека. Когда исследуются случайные выборки людей и ситуаций, индивидуальные различия вносят более весомый вклад в полную дисперсию поведения, чем различия ситуаций. Взаимодействие «человек — ситуация» вносит примерно такой же вклад в полную дисперсию поведения, как и индивидуальные различия, или даже чуть больший. Чтобы идентифицировать общие черты личности, нужно получить оценки индивидуума на множестве ситуаций, используя поддающиеся наблюдению измерения (*dimensions*) и релевантные этим измерениям виды поведения, и совокупить результаты (Epstein, 1980; Kenrick, & Funder, 1988). Несмотря на достигнутое согласие по многим из этих вопросов и тому, что уже известно из исследований, спорных вопросов все же больше, и их число растет в ходе продолжающихся дебатов между сторонниками теории черт и ситуативного подхода (см., например, Funder, 1991).

Конкретный человек. Степень специфичности поведения в различных ситуациях сама варьирует от человека к человеку. Люди различаются по тому, в какой степени они меняют свое поведение в ответ на требования каждой ситуации. В этом отношении умеренная изменчивость (или, если посмотреть под другим углом зрения, непоследовательность, противоречивость) поведения свидетельствует о результативной и адаптивной гибкости, тогда как его чрезмерное постоянство (или последовательность) указывает на неадаптивную ригидность. Кроме того, само множество конкретных ситуаций, на котором поведение человека остается неизменным, может различаться у разных людей. На эту межситуационную устойчивость влияет то, каким образом каждый отдельный человек воспринимает и категоризует ситуации. А это группирование ситуаций, в свою очередь, зависит от целей, мотивов и отношений человека к происходящему, а также от его предыдущего опыта встреч с подобными ситуациями.[2]

---

[1] Ameland, & Borkenau (1986); Bem, & Funder (1978); Endler, & Magnusson (1976); Epstein, (1979, 1980); Epstein, & O'Brien (1985); Hogan, DeSoto, & Solano (1977); Kenrick, & Funder (1988); Mischel (1977, 1979); Mischel, & Peake (1982). См. также краткий обзор дискуссии в Anastasi (1983b).

[2] Такое понимание согласованности поведения ведет начало от идиографического подхода к оценке личности, сформулированного, наряду с другими, Г. Олпортом (1937) и Дж. А. Келли (1963).

Индивидуальные различия в последовательности поведения представляют большой интерес для психологов, вообще говоря, по очевидным причинам. В той мере, в какой эти различия поддаются надежной оценке, их можно было бы использовать в качестве переменной-модератора при прогнозировании поведения. В добавление к этому, с психометрической точки зрения, интраперсональные (*intrapersonal*) и интерперсональные (*interpersonal*) различия в постоянстве (или согласованности) поведения считаются решающими при ослаблении валидности — и надежности — всех психологических мер. Поэтому неудивительно, что в настоящее время предпринимаются попытки, причем с разных направлений, разработать эффективные способы оценки этих различий. Один способ основан на субъективных оценках (*ratings*) людей, которые они дают себе по различным аспектам черты (*trait dimensions*); при этом показатели выводятся из расхождения между такими оценками у каждого человека. Незначительные расхождения между пунктами шкалы отражают такое качество, как *характерность (traitedness)*, которое действительно оказалось связанным с более высокими коэффициентами валидности (Ameland, & Borkenau, 1986; Baumeister, & Tice, 1988). Иной подход нашел воплощение в сформулированном Лэннингом (Lanning, 1991) понятии «шкалируемости» (*scalability*), определяемом как степень, в какой человек сохраняет нормативное упорядочивание поведенческих единиц (*behavioral items*), оцениваемая по данным самоотчета.

Интригующим предложением, которое связывает понятия взаимодействия «человек —ситуация» и социальной желательности через процесс, лежащий в основе ответов на пункты личностного опросника, является сформулированная Джексоном «пороговая теория» реагирования (см., например, Helmes, & Jackson, 1989; Jackson, 1986b). Эта модель базируется на допущении о том, что пункты опросника отображают в миниатюре поведение в реальном мире. Для шкалирования содержания пунктов личностного опросника в данной модели используются методы теории «задание — ответ» (см., главу 7). Джексон полагает, что, так же как ответы многих людей на одно задание можно использовать для получения характеристической кривой задания, ответы одного испытуемого на множество заданий (или пунктов опросника) можно использовать для построения характеристической кривой испытуемого. Такая кривая предсказывала бы вероятность согласия конкретного человека с определенными пунктами опросника и основывалась бы на выраженности фактора желательности для опрашиваемого индивидуума, на его пороге реакции или готовности положительно отвечать на пункты опросника исходя из социальной желательности и на желательности самих пунктов.

Ситуация. Ситуации также различаются по ограничениям, которые они накладывают на поведение. Поэтому-то мы и можем предсказывать на высоком уровне достоверности, что читатели не будут громко разговаривать в библиотеке, а автомобилисты будут останавливаться на красный свет. Аналогично этому, люди — каковы бы ни были черты их личности — по всей вероятности на пляже будут загорать и купаться, а читать в библиотеке. Тем не менее наверняка найдутся такие, кто может проводить свое время, уткнувшись в книгу на пляже, как, впрочем, и такие, кто может потратить слишком много времени, предаваясь мечтам о купании в читальном зале библиотеки. Единственный способ лучше понять ситуационные ограничения поведения — изучить характеристики разнообразных сред, в которых оно реализуется. Недавно обновленный труд Роджера Баркера по экологической психологии снабжает нас многообещаю-

щим инструментарием для классификации условий поведения и описания разнообразных аспектов окружающей среды (Schoggen, 1989).

Кросс-культурные различия можно рассматривать как особый, и более масштабный, случай ситуационной изменчивости. Как таковые, они предоставляют уникальную возможность изучения постоянства (последовательности) и непостоянства (непоследовательности) поведения конкретных людей. С этой целью подходы к кросс-культурному тестированию, рассматриваемые в главах 9 и 12 в контексте мер способностей, могут быть использованы и для исследования других психологических свойств и черт.

Личностные тесты при применении их в культурах, отличных от той, где они были созданы, дают большие расхождения. Любое объяснение таких культурных и субкультурных различий требует конкретного знания условий и обстоятельств, преобладающих в каждой группе. В наибольшей степени эта истина была осознана при оценке психопатологии в несходных культурных группах населения США (Malgady, Rogler, & Costantino, 1987; Paniagua, 1994). Групповые различия по таким тестам, как, например, *MMPI*, часто отражают всего лишь различия в интерпретации отдельных формулировок пунктов или инструкций. Культурные различия в типах поведения, считаемых социально желательными, также могут влиять на показатели тестов. Например, в одних группах повышенные показатели по оценивающей депрессию шкале, возможно, являются следствием сильных традиций самоуничижения и умеренности, тогда как в других они могут указывать на распространенность подлинных эмоциональных проблем, проистекающих из традиционных приемов воспитания детей, конфликтов социальных ролей, фрустраций представителей меньшинств и других широких культурных различий.

Вопрос оценки личности в разных культурах шире проблемы переносимости тестов из культуры в культуру и заключается, скорее, в возможности переноса применяемых для описания поведения концептуальных систем, таких как черты и иерархии черт (Guthrie, Jackson, Astilla, & Elwood, 1983). Вдобавок, так же как раньше в области оценки способностей, сейчас отмечается растущее понимание того, что некоторые важные измерения (*dimensions*) личности не являются универсальными. Соответственно, в дополнение к непрекращающимся попыткам адаптации и перевода англо-американских методик для их применения в других культурах, в настоящее время активно разрабатываются инструменты, предназначенные для оценки специфических измерений (*dimensions*) личности, присущих представителям коренных культурных и субкультурных групп (см., например, Dana, 1993; Lonner, & Berry, 1986). Однако, хотя некоторые конкретные формулировки опросников, да и целые опросники тоже, могут оказаться не переносимыми из одной культуры в другую, имеются веские основания считать, что иерархическая модель черт может быть полезной для интеграции результатов, полученных в разных культурах. То есть если начинать с измерений поведения, определяемого как значимое в каждой конкретной культуре, и проводить их с помощью соответствующих каждой такой культуре инструментов, то появляется возможность идентифицировать психологические конструкты более высокого уровня, которые могут оказаться универсальными или, по крайней мере, распространимыми на достаточно большое число культур (Anastasi, 1992c; Diaz-Guerrero, & Diaz-Loving, 1990).

Черты и ситуации. Наглядной иллюстрацией того, что черты и ситуации отнюдь не являются несовместимыми способами категоризации поведения, служат опросники

для оценки *тестовой тревожности* (I. G. Sarason, 1980). Конкретный пример — Вопросник тестовой тревожности (*Test Anxiety Inventory*[1] [*ТАИ*]), разработанный Спилбергером и его сотрудниками (Spielberger et al., 1980). Этот инструмент, по существу, предназначен для измерения черты. При этом черта определяется на основе строго оговоренного класса ситуаций, относящихся к проведению тестов и экзаменов. Лица с высоким уровнем тестовой тревожности склонны воспринимать оцениваемые ситуации как несущие личную угрозу.

Вопросник состоит из 20 утверждений, описывающих реакции до, во время и после теста или экзамена. Респондентов просят указать, что они *обычно* чувствуют, отмечая, как часто они реагируют описанным в каждом утверждении способом (почти никогда, иногда, часто, почти всегда). К типичным примерам утверждений относятся «Я буквально коченею на важных экзаменах» и «Во время экзаменов я волнуюсь и чувствую себя скованно». *TAI* дает суммарный показатель склонности к тревоге в тестовых ситуациях и частные показатели по двум основным компонентам, установленным посредством факторного анализа, а именно — показатели беспокойства (*worry*) и эмоциональности (*emotionality*). В этом контексте беспокойство определяется как «когнитивная озабоченность последствиями провала», а эмоциональность — как «реакции автономной нервной системы, вызываемые оценочным стрессом» (Spielberger et al., 1980, p. 1).

Еще большая степень ситуационной спецификации предусмотрена в Профиле тестовой тревожности (*Test Anxiety Profile* — Oetting & Deffenbacher, 1980).[2] При работе с эти инструментом респонденты оценивают свои реакции на пункты опросника, покрывающие две области: чувства тревоги и отвлекающие, мешающие мысли. Оба типа показателей тревожности вычисляются отдельно для каждой из шести ситуаций тестирования, в которых, согласно инструкции к опроснику, респонденты должны себя вообразить, и которые варьируют от «теста с множественным выбором ответов» и «опроса без предупреждения» до «свободной беседы».

Возможно, вследствие распространенности боязни испытаний, а значит и тестовой тревожности, и относительной легкости ее исследования в обстановке учебных заведений, теоретическая разработка и эмпирические исследования этой темы продолжаются с неослабевающим энтузиазмом как в США, так и в других странах (Hagtvet, & Johnsen, 1992).

В любом случае, этот конструкт дает хороший образец того, как черты и ситуационные концепты могут принести пользу при категоризации поведения, особенно в сфере личности. В зависимости от назначения теста, конструкты черт могут определяться с различной степенью широты или узости и привязываться к заданным типам ситуаций.

Черты и состояния. Другой способ концептуализации области поведения, оцениваемого личностными тестами, связан с разграничением черт и состояний. Это разграничение наиболее ясно проводится в Опроснике для оценки тревоги/тревожности

---

[1] Маркированный как «Вопросник отношения к тесту» (*Test Attitude Inventory*) на бланке для ответов и указанный под тем же названием в перечнях тестов, опубликованных в 9th *MMY* и *TIP-IV*.

[2] Профиль тестовой тревожности больше не издается. Однако разрешение на его распечатку для исследовательских целей можно получить, обратившись письменно к одному из его создателей, Юджину Иттингу (Eugene Oetting), по следующему адресу: Department of Psychology, Colorado State University, Fort Collins, CO 80523.

(*State-Trait Anxiety Inventory* [*STAI*]), разработанном Спилбергером и его сотрудниками (Spielberger, 1985; Spielberger et al., 1983). При создании этого опросника состояние тревоги (*S-Anxiety*) определялось как временное, преходящее эмоциональное состояние, характеризуемое субъективными чувствами напряжения и опасения (мрачными предчувствиями).

Такие состояния варьируют по интенсивности и флуктуируют во времени. Тревога как состояние (*S-Anxiety*) измеряется с помощью 20 коротких описательных утверждений, ответы на которые даются респондентом в соответствии с тем, как он себя чувствует ***в настоящий момент*** (например, «Я спокоен», «Я встревожен»). Эти ответы фиксируются в форме указания интенсивности переживаемого чувства (вовсе нет; только отчасти; пожалуй, да; совершенно верно).

Тревожность как черта (*T-Anxiety*) относится к относительно устойчивой склонности к тревоге, т. е. к тенденции индивидуума реагировать на ситуации, воспринимаемые как угрожающие, повышением интенсивности тревоги как состояния (*S-Anxiety*). Респондентам дается инструкция указать, что они ***обычно*** чувствуют, отмечая частоту рядом с каждым из 20 применяемых к себе утверждений (почти никогда, иногда, часто, почти всегда). Примеры утверждений: «Я склонен принимать все слишком близко к сердцу» и «Меня ничто не может вывести из равновесия». Люди с высоким уровнем тревожности (*T-Anxiety*) склонны демонстрировать повышение состояния тревоги (*S-Anxiety*) гораздо чаще, чем лица с низким уровнем тревожности, потому что они реагируют на более широкий спектр ситуаций как угрожающих или опасных. Они особенно чувствительны к межличностным ситуациям, представляющим собой некоторую угрозу для их самоуважения, таким как оценка их работы или возможный провал. Однако, будет или нет усиливаться тревога (*S-Anxiety*) в той или иной ситуации, зависит от того, в какой степени конкретный человек воспринимает данную ситуацию как угрожающую или опасную исходя из *своего* прошлого опыта. *STAI* вместе с входящей в комплект его детской версией, Опросником для оценки тревоги/тревожности у детей (*State-Trait Anxiety Inventory for Children* [*STAIC*]), был переведен на 43 языка и диалекта, и по исследованиям с его применением накоплена библиография, включающая более 6000 источников (Spielberger, 1989; Spielberger & Sydeman, 1994).

Дифференциация «состояние/черта» была позднее применена Спилбергером и его коллегами при разработке еще одного инструмента — Опросника для оценки проявлений раздражения и раздражительности (*State-Trait Anger Expression Inventory* [*STAXI*] — Spielberger, 1988; Spielberger, Johnson, Russell, Crane, Jacobs, & Worden, 1985). *STAXI* содержит 44 пункта, отображающих области переживания и проявления раздражения и раздражительности. Область переживания оценивается с помощью двух шкал, аналогичных шкалам *STAI*, а именно: шкалы раздражения (*S-Anger*) и шкалы раздражительности (*T-Anger*), причем последняя, в свою очередь, делится на две подшкалы из четырех пунктов каждая: Раздражительный темперамент (*Anger Temperament*) и Гневная реакция (*Anger Reaction*). Частота проявлений раздражения и раздражительности выявляется с помощью трех шкал, состоящих из восьми пунктов каждая: Внутреннее раздражение (*Anger-in*), Выплескивание раздражения (*Anger-out*) и Сдерживание раздражения (*Anger-Control*). Обзорные материалы по этому тесту (Biskin, 1992; Retzlaff, 1992) можно найти в 11-м выпуске Ежегодника психических измерений (*11th MMY*).

Современное состояние личностных опросников

Вдобавок к общим проблемам, встречающимся во всех областях психологического тестирования, разработчикам и пользователям личностных опросников приходится сталкиваться с дополнительными трудностями. Вопрос сознательного обмана при оценивании личности стоит гораздо острее, чем при тестировании способностей. Измеряемое личностными тестами поведение имеет более выраженную временную динамику по сравнению с поведением, измеряемым тестами способностей. Последний факт усложняет определение надежности теста, так как случайные временные флуктуации в выполнении теста, по всей видимости, смешиваются с общими, закономерными изменениями поведения. Даже при тестировании через относительно короткие интервалы времени не приходится ожидать, что вариация тестовых ответов ограничивается только тестом и не характеризует область нетестируемого поведения. Сюда же можно отнести проблему большей ситуационной специфичности ответов тестируемых в некогнитивной области по сравнению с когнитивной.

В 1990-х гг. произошло оживление исследований, не уходящих от сложностей оценивания личности, а, напротив, нацеленных на поиски новаторских решений этих старых проблем. Данный период характеризуется существенными теоретическими и методологическими прорывами.[1] Предшествовавшая ему критика измерений личности безусловно принесла пользу и отчасти стимулировала последующие разработки в этой области психометрии. Однако в своем рвении искоренить ошибочные представления мы должны сохранять бдительность, чтобы не расстаться заодно с правильными и полезными понятиями. Например, за звучащим время от времени предложением полностью отказаться от диагностического тестирования личности и понятия черты скрывается неоправданно узкое определение этих терминов. Диагноз совсем не обязательно предполагает навешивание ярлыков на людей, использование традиционных психиатрических категорий или применение медицинской модели «болезни». К диагностическому тестированию следует прибегать как к вспомогательному средству при описании и объяснении поведения индивидуума, выявлении его проблем и выработке решений относительно дальнейших действий. Точно так же и черты личности относятся к таким категориям, в которых с необходимостью приходится описывать поведение независимо от контекста, научного или какого-то другого, разумеется, если мы хотим хоть что-то понять в поведении конкретного человека. Оптимальная широта таких категорий будет меняться в зависимости от конкретных целей оценивания. Представленные в главе 11 модели иерархической организации черт с равным успехом могут использоваться и при описании некогнитивного поведения. При одних обстоятельствах лучше всего подойдут относительно широкие, общие черты, а при других — необходимо будет оценить узкие, точно определенные формы поведения.

---

[1] Работы Брутона (Broughton), применяющего понятие прототипа и методы многомерного шкалирования для оценки личности, могут служить примером ряда наиболее интересных и быстро развивающихся методологических разработок в данной области (Broughton, 1990; Broughton, Boyes, & Mitchell, 1993). Что касается теории, то здесь круговая модель (*circumplex model*) обеспечила основу для интеграции различных традиций в сфере межличностного поведения (см., например, Hofstee et al., 1992; J. S. Wiggins, 1996; Wiggins, & Pincus, 1992). В добавление к этому предметом обсуждения и исследования все чаще становятся многочисленные взаимосвязи между областями личности и интеллекта (см. главу 16).

14 ИЗМЕРЕНИЕ ИНТЕРЕСОВ И АТТИТЮДОВ[1]

Характер и сила интересов и аттитюдов являются важным аспектом личности. Эти характеристики человека существенно влияют на его учебные и профессиональные достижения, отношения с другими людьми, на удовольствие, получаемое от занятий в свободное время, и на другие важные стороны его повседневной жизни. Хотя тесты типично направлены на измерение той или другой из этих переменных, имеющиеся инструменты невозможно строго классифицировать в соответствии с такими дискретными категориями, как интересы и аттитюды. Частичное перекрытие является в этой области правилом. Так, о вопроснике, предназначенном для оценки относительной силы интересов к занятиям научными исследованиями, искусством или традиционной работой, можно, как правило, сказать, что он измеряет и аттитюды индивидуума в отношении чистой науки, искусства ради искусства, практической работы и т. д.

Ценности также явно связаны с выбором образа жизни и часто рассматриваются в одной связке с интересами, аттитюдами и предпочтениями. Специалистами в области социальной психологии и психологии личности проведено, да и сейчас проводится, большое количество исследований по проблеме ценностей, включая ряд интересных проектов по изучению универсальности ценностей в широком диапазоне культур (S. H. Schwartz, 1992, 1994; S. H. Schwartz & Sagiv, 1995; Super, & Sverko, 1995). В последние годы, однако, появилось относительно мало разработок в области стандартизованных, доступных для приобретения инструментов, единственное назначение которых — оценка ценностей. Это связано с рядом проблем, характерных для измерения ценностей, прежде всего трудностей систематического — и на подходящем уровне абстракции — отбора образцов из областей ценностей. Вдобавок ко всему не-

---

[1] Англоязычным термином *attitude* обозначается выражаемое в вербальной или невербальной форме отношение человека к различным аспектам действительности, включая и самого себя. Можно согласиться с В. С. Магуном, что такие варианты перевода этого термина на русский язык, как «установка», «социальная установка» и «отношение», не вполне адекватны. Поэтому в рамках этой книги везде, где слово *attitude* несет терминологическую нагрузку, оно передается лексической калькой «аттитюд». — *Примеч. науч. ред.*

которые из ранних и наиболее широко используемых средств измерения ценностей оказались несовместимыми с тем способом, каким по прошествии времени стали концептуализировать ценности в данной области (Braithwaite & Scott, 1991).[1] Хотя некоторые «независимые» тесты наподобие Инвентаря жизненных ценностей (*Life Values Inventory* — Brown, & Crace, 1996) и Шкалы ценностей (*The Values Scale* — Nevill, & Super, 1989) все еще продолжают издаваться и перерабатываться, формальные (строгие) средства оценки ценностей по большей части включаются теперь в состав инструментов, предназначенных для облегчения принятия решений, связанных с выбором или сменой профессии, и для оценки относящихся к работе аттитюдов и мотивов.[2]

Изучение *интересов* получило сильнейший импульс к развитию из сферы консультирования по вопросам образования и выбора профессии. В несколько меньшей степени разработка тестов в этой области стимулировалась также задачами профессионального отбора и распределения персонала. С позиций как работника, так и его нанимателя учет интересов конкретного человека имеет практическое значение. Оценка *мнений и аттитюдов* первоначально была преимущественно проблемой социальной психологии. Аттитюды в отношении непохожих («чужих») групп, например, имеют очевидные последствия для межгрупповых отношений. Аналогично этому, точное измерение и прогнозирование общественного мнения по широкому кругу вопросов представляет одинаково глубокий интерес как для социального психолога, так и для бизнесмена, политика и работников других сфер деятельности. Кроме того, измерение мнений и аттитюдов делает большие успехи в таких областях, как исследования потребителя и взаимоотношений между членами трудового коллектива.

Все кратко рассматриваемые в этой главе инструменты представляют собой опросники (инвентари) типа стандартизованных самоотчетов, разработанные с применением одного или нескольких методов, описанных в главе 13. Следует, однако, отметить, что в этой области, как и вообще в сфере измерения конструктов личности, непрерывно развиваются и опробуются другие подходы и методы, которые будут обсуждаться в главах 15 и 16.

Инвентари интересов: текущее состояние

Подавляющее большинство инвентарей интересов[3] предназначено для оценки интересов индивидуума в разных областях труда. Некоторые из этих инструментов обеспечивают также анализ интересов к образовательным программам и курсам наук, которые, в свою очередь, увязываются с выбором карьеры. Хотя частота применения

---

[1] Например, один из первых инвентарей ценностей — «Изучение ценностей» (*Study of Values* — Allport, Vernon, & Lindzey, 1960) — состоял большей частью из пунктов, относящихся к предпочтениям, мнениям и интересам, а не к ценностям как таковым (в настоящее время этот инструмент не издается).

[2] Описания многих инструментов, предназначенных преимущественно или исключительно для оценки ценностей, можно найти в Anastasi (1988b, p. 580–583), Braithwaite, & Scott (1991), Dawis (1991, p. 845–850).

[3] В настоящее время вышло уже третье издание полного руководства по инвентарям интересов и другим инструментам оценки карьеры, подготовленное Национальной ассоциацией карьерного роста (*National Career Development Association* — Kapes, Mastie, & Whitfield, 1994). Что касается обзора важных теоретических и методологических вопросов в этой области, а также примеров ряда основных измерительных инструментов, см. Borgen (1986) и Hansen (1990).

тестов в консультировании практически не меняется с 1950-х гг., использование тестов интересов заметно возросло относительно использования тестов способностей (Zytowski, & Warman, 1982). Более современные инвентари интересов (новые или переработанные) отражают важные перемены в профконсультировании. Одна из таких перемен имеет отношение к увеличивающемуся значению *самоизучения (self-exploration)*. Все больше инструментов предоставляют каждому человеку возможность изучить подробные результаты теста и соотнести их с информацией о профессиях и другими данными о личных качествах и собственном опыте. Практическому обучению принимать решения в отношении своей карьеры также уделяется растущее внимание. Обсуждаемые в этой главе инвентари профессиональных интересов следует рассматривать с учетом того, что происходит в области оценки карьеры (см. главу 17).

Вторая перемена, связанная с первой, касается цели измерения интересов. В наши дни все большее значение придается *расширению вариантов выбора карьеры (expanding the career options)*, доступной для конкретного человека. Фактически, термин «поисковая валидность» (*exploration validity*) используется для обозначения того воздействия, которое инвентари интересов могут оказывать на человека через активизацию и расширение поведения, способствующего ознакомлению с профессиями и выбору карьеры (см., например, Randahl, Hansen, & Haverkamp, 1993). Таким образом, инвентари интересов, так же как и более сложные программы профориентации, упоминаемые в главе 17, применяются для ознакомления индивидуума с подходящими для него профессиями, которые в противном случае могли оказаться для него неизвестными.

Третья важная перемена уже напрямую связана с этим расширением вариантов выбора карьеры. Она касается усилий по *устранению половой дискриминации (sex fairness)* в инвентарях интересов. В общем, такие инвентари сравнивают выраженные интересы индивидуума с интересами, типичными для людей, занятых в различных профессиях. Это делается или на этапе количественной обработки ответов респондента на пункты опросника, или на этапе интерпретации показателей по широким областям интересов, или же на обоих этапах. Хотя этот подход безусловно представляет собой объективную, эмпирическую процедуру для оценивания интересов человека, он несет в себе тенденцию к закреплению существующих групповых различий в профессиях. Если существует большой разрыв в относительной представленности мужчин и женщин в ряде профессий, как, например, это имеет место в области техники или сестринского дела, такие различия так или иначе будут сказываться на интерпретации результатов, полученных представителями мужского и женского пола по инвентарям интересов. По этой причине способам снижения возможной необъективности инвентарей интересов, связанной с полом респондентов, было посвящено немало дискуссий и тщательных исследований (Tittle, & Zytowski, 1978; Zytowski, & Borgen, 1983). Кроме того, был подготовлен и широко распространяется комплект методических указаний по оцениванию систематической ошибки по полу и отсутствия половой дискриминации в инвентарях интересов, используемых в профориентации и профконсультировании.[1] Почти каждый вновь созданный или пересмотренный инвентарь обнаруживает влияние этих указаний и включает некоторые меры для обеспечения

[1] Комплект подготовлен в рамках исследования, проведенного Национальным институтом образования (*National Institute of Education*). Его копия приведена в работе Tittle, & Zytowski (1978, p. 151–153).

беспристрастности к респондентам разного пола. Одним из предпринятых в этом направлении шагов стало устранение смещения по полу в словесных формулировках пунктов инвентарей. В добавление к этому, другие общие решения касались уравновешивания содержания пунктов исходя из типичных различий в полоролевой социализации и обеспечения максимально объективных норм для каждой половой группы по шкалам инвентаря. Эти меры повысили качество исходных данных, поступающих в распоряжение профконсультантов и других пользователей инвентарей интересов. Однако из-за сохраняющихся половых различий в структуре интересов и в составе профессиональных групп, остаются и острые вопросы в области социальной политики, затрагивающие интерпретацию и использование данных, получаемых с помощью инвентарей интересов, требующие внимательного отношения при оценке карьеры женщин (Hackett, & Lonborg, 1994).

С недавних пор на первый план выступила другая, хотя и близкая к только что обсуждавшейся, задача, возникшая в ответ на принятие закона, цель которого — обеспечить профориентационную работу среди инвалидов и иных групп населения, оказавшихся в неблагоприятных условиях[1] (L. S. Gottfredson, 1986b; Reed, Rotatori, & Day, 1990; Szymula, 1990). Среди прочего, особые потребности инвалидов в том, что касается профессионального оценивания, были учтены в разработке ряда рисуночных инвентарей интересов, в которых используются картинки, диафильмы или видеокассеты (Elksnin, & Elksnin, 1993; Kapes et al., 1994, p. 307–345). Первое поколение этих инструментов, сконструированных таким образом, чтобы обойти требования традиционных бланковых тестов интересов к навыкам чтения, оказались несостоятельными с психометрической точки зрения, особенно в области валидности. Тем не менее рисуночные инвентари интересов отображают как новаторский способ предъявления стимульного материала такого рода тестов, который, вероятно, будет и дальше развиваться по мере совершенствования технологии, так и потребность в расширении ассортимента альтернативных средств измерения для специфических популяций. Модификации традиционных инструментов, наподобие описанной в главе 9 в связи тестами способностей для специфических популяций, могут, разумеется, в равной мере использоваться и при оценивании интересов. Закон об инвалидах-американцах от 1990 г. (*Americans with Disabilities Act of 1990* [P. L. 101–336]) безусловно будет стимулировать интерес и дальнейшую работу в этой области (Bruyere & O'Keeffe, 1994).

Инвентарь интересов Стронга (Strong Interest Inventory™ — SII)

Истоки и развитие *SII*. Этот инвентарь интересов, последняя редакция которого была опубликована в 1994 г., имеет длинную историю. Общий подход к его разработке был сформулирован Э. К. Стронгом-младшим (E. K. Strong, Jr.) в 1919/20 учебном году на семинаре по изучению интересов для студентов-дипломников Технологического института Карнеги (D. P. Campbell, 1971, chap. 11; Fryer, 1931, chap. 3). После

[1] Два конкретных примера таких законов — Акт Карла Д. Перкинса о профессиональном образовании от 1984 г. (*Carl D. Perkins Vocational Education Act of 1984* [P. L. 98–524]) и Поправки 1990 г. к Акту Карла Д. Перкинса о профессиональном и ремесленном образовании (*Carl D. Perkins Vocational and Applied Technology Education Act Amendments of 1990* [P. L. 101–392]).

того как в 1927 г. был впервые опубликован Бланк профессиональных интересов Стронга (*Strong Vocational Interest Blank®* [*SVIB*]), психометрия обогатилась двумя принципиально новыми методическими приемами измерения профессиональных интересов. Во-первых, пункты опросника предполагали дихотомическую реакцию — «нравится/не нравится» — в отношении широкого множества конкретных занятий, предметов или типов людей, с которыми респондент обычно сталкивается в повседневной жизни. Во-вторых, эти реакции респондента эмпирически приводились в соответствие с различными профессиями. Таким образом, эти инвентари интересов оказались в числе первых тестов, в которых использовался метод привязки заданий к эмпирическому критерию, впоследствии примененный при разработке таких личностных опросников, как *MMPI* и *CPI* (глава 13). Было обнаружено, что лица, занятые в ряде профессий, характеризуются общими интересами, которые отличают их от лиц других профессий. Эти различия в интересах распространялись не только на вопросы, имеющие непосредственное отношение к профессиональной деятельности, но и на школьные предметы, хобби, любимые виды спорта, театральные постановки или книги, социальные отношения и многие другие стороны обыденной жизни. Поэтому оказалось возможным составить опросник, который бы позволял выявлять интересы индивидуума к знакомым вещам и таким путем определять, насколько близко его интересы сходятся с интересами людей, успешно занимающихся определенной профессиональной деятельностью.

Начиная с 1970-х гг., в последовательных редакциях *SII* было реализовано множество нововведений (D. P. Campbell, 1974; D. P. Campbell, & Hansen, 1981; Hansen, & D. P. Campbell, 1985; Harmon, Hansen, Borgen, & Hammer, 1994). К числу главных изменений относятся: а) включение теоретической оболочки, направляющей организацию и интерпретацию показателей; б) объединение более ранних мужских и женских форм опросника и ренормирование всех профессиональных шкал на новых выборках представителей мужского и женского пола; в) существенное увеличение числа шкал для профессий, освоение которых не требует окончания колледжа, — эта категория профессий была слабо представлена в более ранних редакциях *SII*.

***SII*–Форма Т317: общая характеристика.** Современный инвентарь интересов Стронга состоит из 317 пунктов, сгруппированных в восемь разделов. В первых пяти разделах респондент отмечает свои предпочтения начальными буквами слов *Like* (Нравиться), *Indifferent* (Безразлично) или *Dislike* (Не нравится). Пункты в этих пяти разделах относятся к следующим категориям: профессии, школьные предметы, занятия (например, публичное выступление, починка часов, сбор денег на благотворительные цели), развлечения и повседневное общение с различного типа людьми (например, дряхлыми стариками, армейскими офицерами, людьми, рискующими жизнью). Два дополнительных раздела требуют от респондента отдать предпочтение одному из пары занятий (например, иметь дело с вещами или иметь дело с людьми) и одному из всех возможных парных сочетаний четырех объектов из мира труда: идей, фактов, вещей и людей. Наконец, часть реакций респондента касается самоописаний, просматривая набор которых, он должен пометить один из трех вариантов ответа: «да», «нет» или «?».

Инвентарь Стронга обрабатывается только на компьютере, либо в специальных центрах по обработке *SII*, указываемых издателями, либо с помощью программного обеспечения, которое можно приобрести у тех же издателей в различных вариантах.

На рис. 14–1 показана первая страница самого свежего профиля Стронга, которая дает «моментальный снимок» (*snapshot*), или краткую сводку наивысших показателей тестируемого по основным шкалам инвентаря.[1] В инвентаре Стронга есть три уровня показателей, различающихся по широте. Самыми широкими и наиболее полными являются показатели шести Общих профессиональных тем (*General Occupational Themes*); следующий уровень включает 25 шкал Основных интересов (*Basic Interest Scales*) и, наконец, самый конкретный уровень представлен 211 шкалами Профессий (*Occupational Scales*), доступных для выбора в настоящее время. В добавление к этим шкалам Форма *T317 SII* предусматривает получение показателей по четырем новым шкалам Личного стиля (*Personal Style Scales*), которые оценивают предпочтения в таких областях, как Стиль работы (*Work Style*), Среда обучения (*Learning Environment*), Стиль руководства (*Leadership Style*) и Рискованные действия/Авантюры (*Risk Taking/Adventure*). Кроме того, данная форма профиля дает набор Административных индексов (*Administrative Indexes*), включая суммарное число ответов (для обнаружения чрезмерного количества пропусков), число редких или необычных ответов и процент каждого из трех вариантов ответов для каждого из восьми разделов инвентаря Стронга. Эти индексы можно использовать как средства контроля недобросовестности или специфических установок на ответ.

В основу используемой в *SII* классификации профессиональных интересов положена теоретическая модель, разработанная Джоном Холландом (J. Holland, 1966, 1985/1992) и подтвержденная обширными исследованиями как самого Холланда, так и других независимых исследователей. *Общие профессиональные темы*, определяемые моделью Холланда, имеют следующие названия: Реалистическая [Р] (*Realistic* [*R*]), Исследовательская [И] (*Investigative* [*I*]), Художественная [X] (*Artistic* [*A*]), Социальная [С] (*Social* [*S*]), Предпринимательская [П] (*Enterprising* [*E*]) и Конвенциальная [К] (*Conventional* [*C*]).[2] Каждая тема характеризует не только тип человека, но и тип рабочей среды, которую такой человек находит наиболее благоприятной. Каждая из таких сред имеет тенденцию заполняться людьми соответствующего типа, и именно они, в целом, преобладают и занимают господствующее положение в подходящей для себя среде. Согласно Холланду, конкретные люди не распределяются строго по каждому из шести главных типов; скорее, они характеризуются степенью сходства с одним или более типами. Такие сочетания типов, упорядоченные по степени сходства, обеспечивают разнообразие паттернов или «кодов» для писания широкого многообразия индивидуальных различий.

На рис. 14–2 шесть Общих профессиональных тем представлены в виде углов разработанной Холландом шестиугольной модели. В целях краткости каждую тему принято обозначать первой буквой ее названия. Таким образом, порядок расположения тем в углах шестиугольника образует акроним *R-I-A-S-E-C* (в переводе на русский —

---

[1] Полный профиль SII был переконструирован и теперь состоит из шести страниц. На лицевой стороне каждой страницы приводятся показатели тестируемого по всем шкалам, а на оборотной стороне — основная информация о значении показателей и предложения о том, как приступить к поиску профессии. В добавление к этому профилю можно получить более объемное и носящее описательный характер заключение, содержащее качественную интерпретацию показателей и специально приспособленное к уровню и запросам конкретного респондента.

[2] Эти темы в значительной степени пересекаются с оценочными аттитюдами, измеряемыми с помощью опросника «Изучение ценностей» (Allport et al., 1960), формулировки которых, в свою очередь, были подсказаны работой Э. Шпрангера «Человеческие типы» (Types of Men, 1928).

STRONG INTEREST INVENTORY

Profile report for: CLIENT 1
ID:
Age: 20
Gender: Male

Date tested:
Date scored:

Page 1 of 6

SNAPSHOT: A SUMMARY OF RESULTS FOR CLIENT 1

GENERAL OCCUPATIONAL THEMES

The General Occupational Themes describe interests in six very broad areas, including interest in work and leisure activities, kinds of people, and work settings. Your interests in each area are shown at the right in rank order. Note that each Theme has a code, represented by the first letter of the Theme name.

You can use your Theme code, printed below your results, to identify school subjects, part-time jobs, college majors, leisure activities, or careers that you might find interesting. See the back of this Profile for suggestions on how to use your Theme code.

THEME CODE	THEME	VERY LITTLE INTEREST	LITTLE INTEREST	AVERAGE INTEREST	HIGH INTEREST	VERY HIGH INTEREST	TYPICAL INTERESTS
I	INVESTIGATIVE			✓			Researching, analyzing
R	REALISTIC			✓			Building, repairing
C	CONVENTIONAL			✓			Accounting, processing data
A	ARTISTIC			✓			Creating or enjoying art
S	SOCIAL		✓				Helping, instructing
E	ENTERPRISING		✓				Selling, managing

Your Theme code is IRC—(see explanation at left).
You might explore occupations with codes that contain any combination of these letters.

BASIC INTEREST SCALES

The Basic Interest Scales measure your interests in 25 specific areas or activities. Only those 5 areas in which you show the *most* interest are listed at the right in rank order. Your results on all 25 Basic Interest Scales are found on page 2.

To the left of each scale is a letter that shows which of the six General Occupational Themes this activity is most closely related to. These codes can help you to identify other activities that you may enjoy.

THEME CODE	BASIC INTERESTS	VERY LITTLE INTEREST	LITTLE INTEREST	AVERAGE INTEREST	HIGH INTEREST	VERY HIGH INTEREST	TYPICAL ACTIVITIES
R	ATHLETICS					✓	Playing or watching sports
I	MATHEMATICS					✓	Working with numbers or statistics
I	MEDICAL SCIENCE				✓		Working in medicine or biology
R	MECHANICAL ACTIVITIES			✓			Working with tools and equipment
A	APPLIED ARTS			✓			Producing or enjoying visual art

OCCUPATIONAL SCALES

The Occupational Scales measure how similar your interests are to the interests of people who are satisfied working in those occupations. Only the 10 scales on which your interests are *most* similar to those of these people are listed at the right in rank order. Your results on all 211 of the Occupational Scales are found on pages 3, 4, and 5.

The letters to the left of each scale identify the Theme or Themes that most closely describe the interests of people working in that occupation. You can use these letters to find additional, related occupations that you might find interesting. After reviewing your results on all six pages of this Profile, see the back of page 5 for tips on finding other occupations in the Theme or Themes that interest you the most.

THEME CODE	OCCUPATION	VERY DISSIMILAR	DISSIMILAR	MID-RANGE	SIMILAR	VERY SIMILAR
RI	ENGINEER					✓
RI	RADIOLOGIC TECHNOLOGIST					✓
IR	DENTIST				✓	
R	PLUMBER				✓	
CI	ACTUARY				✓	
IR	OPTOMETRIST				✓	
CE	ACCOUNTANT				✓	
IR	CHEMIST				✓	
R	AUTO MECHANIC				✓	
RIC	ELECTRICIAN				✓	

***PERSONAL STYLE SCALES** measure your levels of comfort regarding Work Style, Learning Environment, Leadership Style, and Risk Taking/Adventure. This information may help you make decisions about particular work environments, educational settings, and types of activities you would find satisfying. Your results on these four scales are on page 6.*

CPP CONSULTING PSYCHOLOGISTS PRESS, INC. • 3803 Bayshore Road, Palo Alto, CA 94303

Рис. 14–1. Краткая сводка результатов по Инвентарю интересов Стронга. Первая страница профиля дает «моментальную фотографию» показателей 20-летнего студента колледжа, выбирающего специализацию

(Форма перепечатана с разрешения издательства из Strong Interest Inventory: Applications and technical guide, *by Lenore W. Harmon, Jo-Ida C. Hansen, Fred H. Borgen, and Allen L. Hammer (p. 236). Copyright © 1994 by Stanford University Press)*

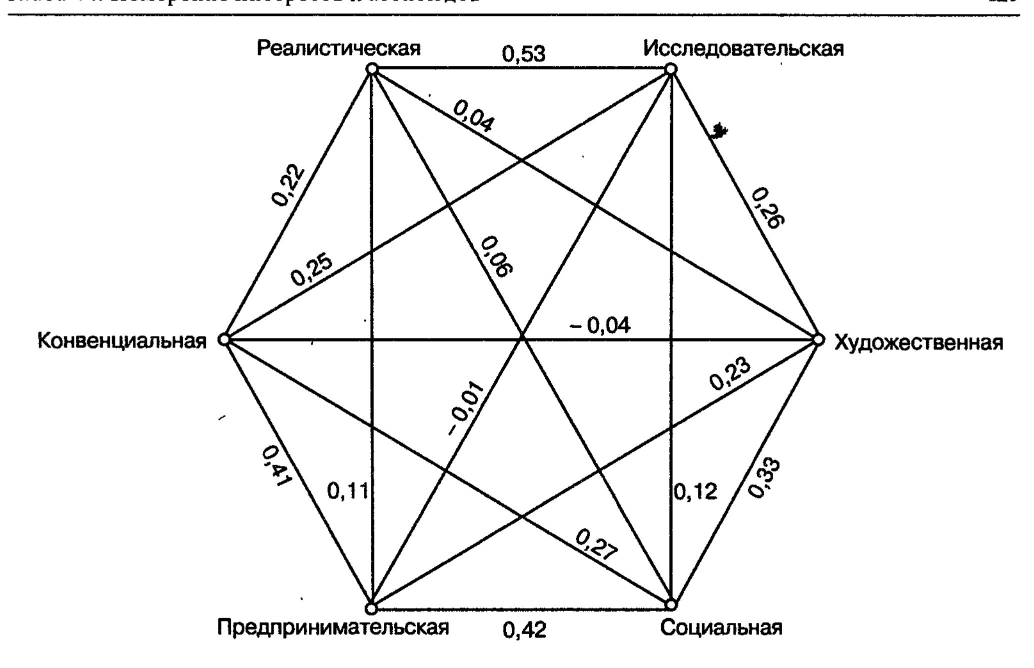


Рис. 14–2. Шестиугольная модель Общих профессиональных тем Холланда. Корреляции для каждой пары тем вычислены на основе Генеральной эталонной выборки (*General Reference Sample* [*GRS*]), состоящей из 9484 мужчин и 9467 женщин

(Перепечатано с разрешения издательства из Strong Interest Inventory: Applications and technical guide, *by Lenore W. Harmon, Jo-Ida C. Hansen, Fred H. Borgen, and Allen L. Hammer (p. 51). Copyright © 1994 by Stanford University Press)*

Р-И-Х-С-П-К), в свою очередь используемый для краткого обозначения этой модели. Следует отметить, что самые высокие корреляции были получены между шкалами тем, занимающих соседние позиции по периметру шестиугольника. Самые низкие корреляции обнаружились между шкалами тем, находящихся на противоположных концах диагоналей. Например, корреляция Реалистической шкалы с Исследовательской шкалой равна 0,53, тогда как корреляции той же шкалы с Художественной и Социальной шкалами составляют, соответственно, всего лишь 0,04 и 0,06. Точно так же, если шкалы Профессий нанести в виде точек на шестиугольник Холланда, большая их часть разместится по его периметру в ожидаемом порядке. Например, профессия инженера-конструктора, помеченная кодом РИ (*RI*), попадет между Реалистической и Исследовательской темами, а профессия банкира (код КП (*CE*)) — между Конвенциальной и Предпринимательской темами. Обычно профессии с высоким показателем по одной теме имеют низкий показатель по противоположной (расположенной на другом конце диагонали) теме (примером могут служить Художественная и Конвенциальная темы). В тех случаях, когда какая-то шкала Профессий обнаруживает существенные корреляции с темами, размещенными в противоположных от нее точках шестиугольника, соответствующая профессиональная группа часто состоит из разнородных подгрупп с заметно различающимися профессиональными функциями.

Двадцать пять шкал *Основных интересов* классифицируются в соответствии с шестью Общими профессиональными темами. Эти шкалы состоят из кластеров взаимно

коррелированных пунктов. Шкалы Основных интересов являются более однородными по содержанию, чем шкалы Профессий, и потому могут облегчить понимание причин того, что человек имеет высокие показатели по определенной шкале Профессий.

Шкалы *Профессий*, составлявшие основу оригинального *SVIB*, теперь сгруппированы по соответствующим Общим профессиональным темам. В ходе многолетних, продолжающихся и по сей день, исследований инвентаря Стронга в его состав добавлялись новые шкалы, а старые обновлялись на основе свежих критериальных выборок. Форма *T317* включает в себя 211 шкал Профессий, 83 % которых нормированы на выборках, полученных в 1980-е и 1990-е гг. За исключением семи шкал, для всех остальных оказалось возможным скомплектовать группы респондентов в количестве, достаточном для разработки шкал, нормированных отдельно для мужчин и женщин, что дало 204 шкалы, по 102 для каждого пола. До сих пор 5 шкал имеют нормы только для женщин, а 2 — только для мужчин.

Большинство выборок, использованных при разработке каждой шкалы Профессий, состоит из 200 или более человек, хотя фактическое количество членов таких выборок колеблется от 60 до 1187 человек. Для проводившегося в 1994 г. пересмотра *SII* было протестировано более 55 000 человек, и чуть меньше 40 000 из них полностью соответствовали требованиям, позволяющим использовать полученные данные для конструирования шкал. Профессиональные критериальные группы состояли преимущественно из лиц в возрасте от 25 до 60 лет, имевших как минимум трехлетний стаж работы по избранной профессии, удовлетворенных (по их словам) своей работой и выполнявших типичные для представителей своей профессиональной группы обязанности. Есть еще две Генеральных эталонных выборки (ГЭВ)[1], включающие 9484 мужчин и 9467 женщин, протестированных в 1990-х гг. ГЭВ охватывают 98 различных профессий, из которых 90 представлены 200 членами, случайно выбранными из наличной совокупности представителей каждой профессии. Восемь профессий представлены выборками, объем которых колебался от 92 до 195 человек.[2] Уровень образования входящих в ГЭВ для *SII* лиц гораздо выше уровня образования населения США в целом: фактически, 80 % членов ГЭВ имеют степени бакалавра, магистра или даже еще более высокие ученые степени (Harmon et al., 1994, p. 110). Эта доля выпускников колледжей и университетов существенно превышает сопоставимые цифры в генеральной совокупности (т. е. в структуре всего населения США) даже для занятых умственным трудом лиц, которые составляют подавляющее большинство в ГЭВ для *SII*. На этом основании самую последнюю редакцию инвентаря Стронга, как его более ранние версии, можно критиковать за недостаточную репрезентативность выборок (Worthen, 1995).

Для распределения пунктов SII по шкалам применялись два разных метода. Пункты в шкалах Общих профессиональных тем и Основных интересов группировались в однородные кластеры на основе сходства их содержания и способа ответов на них респондентов, которое оценивалось по результатам факторного анализа. С другой стороны, пункты для каждой шкалы Профессий отбирались и взвешивались на основе процентных *различий* ответов на каждый пункт между критериальной профессиональной группой и двумя разнополыми ГЭВ. В шкале для бухгалтеров-женщин, на-

---

[1] Другой возможный перевод — Общие контрольные выборки (ОКВ). — *Примеч. науч. ред.*

[2] Полный перечень всех профессиональных выборок, использовавшихся при пересмотре *SII* в 1994 г., можно найти в Harmon et al., (1994, Appendix A).

пример, весовой коэффициент +1 показывает, что ответ встречается чаще, а весовой коэффициент –1 указывает на то, что ответ встречается реже в выборке бухгалтеров-женщин, чем у женщин вообще. Ответы, по которым не удается достаточно четко разграничить женщин-бухгалтеров и женщин из ГЭВ, не засчитываются по шкале бухгалтеров-женщин, независимо от того, насколько часто такие ответы выбирались бухгалтерами. Итоговый первичный показатель по каждой шкале Профессий представляет собой алгебраическую сумму положительных и отрицательных весов ответов респондента.

Обработка и интерпретация результатов. Все показатели по инвентарю Стронга представляются в виде стандартных показателей со средним, равным 50, и $SD = 10$. Для шкал Общих профессиональных тем и Основных интересов нормативной выборкой, на основе которой рассчитываются стандартные показатели, служит объединенная по полу ГЭВ ($N = 18\,951$). Однако и интерпретационные категории, и графические отображения показателей на профилях основываются на нормах того пола, к которому принадлежит тестируемый. Сравнения с нормами для другого пола также возможны и проводятся относительно данных, графически представленных на соответствующих формах профиля. По шкалам Профессий каждый респондент в действительности получает два стандартных показателя — один, выведенный на основе женской профессиональной выборки, а другой — на основе мужской профессиональной выборки. Эти процедуры составления заключения по тесту хотя и нацелены преимущественно на сравнения лиц одного пола, предоставляют в распоряжение консультантов и респондентов данные, позволяющие проводить межполовые сравнения для более полной и более действенной интерпретации паттернов ответов в индивидуальных случаях.

Техническое руководство по SII и его практическим приложениям (*Strong Interest Inventory: Applications and technical guide* — Harmon ét al., 1994) отличается высоким качеством и содержит множество сведений, помогающих консультанту проинтерпретировать результаты и обсудить их возможные последствия с клиентом. Например, профили, отличающиеся внутренней согласованностью результатов, обеспечивают, в целом, более высокую предсказуемость. Впрочем, некоторая несогласованность результатов, скажем, между шкалами Основных интересов и шкалами Профессий, может оказаться полезной для понимания природы и источников выраженных предпочтений респондента. Кроме того, результаты, полученные с помощью инвентаря Стронга можно распространить на другие, родственные профессии благодаря связям, установленным с профессиями, перечисленными в составленном министерством труда США (1991) *Словаре названий профессий (Dictionary of Occupational Titles [DOT])* и в других аналогичных справочных материалах (Gottfredson, & Holland, 1989; Harmon et al., 1994, Appendixes A & B). Техническое руководство по *SII* включает также несколько глав, посвященных использованию инвентаря Стронга в работе со специфическими популяциями, такими как представители других культур и инвалиды. В дополнение к этому руководству есть еще несколько публикаций, назначение которых — помочь клиентам и консультантам в понимании и использовании результатов *SII* (см., например, Borgen, & Grutter, 1995; Hirsh, 1995; Prince, 1995).

Психометрическая оценка. Инвентарь Стронга является предметом непрекращающихся исследований, что позволило получить обширные данные по его надежности и валидности (D. P. Campbell, 1971, 1977; Hansen & Campbell, 1985; Harmon et al., 1994).

Что касается шкал Профессий, медианный коэффициент ретестовой надежности, определяемой для трех- и шестимесячного интервалов на выборке из 191 работающего по найму взрослого, составил 0,90; значения аналогичных коэффициентов надежности для шкал Основных интересов и Общих профессиональных тем равны соответственно 0,86 и 0,89. Долговременная стабильность шкал Профессий из предыдущих редакций инвентаря Стронга также весьма высока. Корреляции для периодов продолжительностью до 20 лет имеют по большей части величины порядка 0,60–0,70 для лиц младше 25 лет и порядка 0,80 для лиц старше 25 лет.

Данные, используемые для текущей валидизации инвентаря Стронга, касаются степени дифференциации между разными профессиональными выборками, а также между профессиональными и контрольными (эталонными) выборками. Для 211 профессиональных выборок, включенных в пересмотр *SII* 1994 г., медианное перекрытие составило 36 %, что свидетельствует о среднем расхождении шкал Профессий и ГЭВ на величину чуть меньше двух стандартных отклонений. Прогностическая валидность проверялась на нескольких выборках в рамках длительных периодов для предыдущих редакций инвентаря Стронга. Данные говорят о наличии значительного соответствия между начальным профессиональным профилем и выбранной в конечном итоге профессией. Характерным примером может служить проведенное через 40 лет повторное изучение выборки психологов, профессиональная карьера которых выявила ряд продуктивных связей между сглаженностью/выраженностью их первичных профилей и такими событиями, как частота смены работы и переход с преподавательской или исследовательской работы на административную или прикладную (Vinitsky, 1973). В другом исследовании было установлено поразительно высокое межкультурное сходство профилей в выборках психологов, протестированных в 9 западных странах (Lonner, & Adams, 1972). Несмотря на широкую исследовательскую базу, созданную за многие годы работы с инвентарями Стронга, и на их очевидную основательность как индикаторов профессионального выбора и срока работы по избранной профессии, необходимо предпринять новые исследования самой свежей версии *SII* для получения прямых данных о прогностической валидности этого инструмента.

Что касается валидизации конструктов, положенных в основу *SII*, особое значение здесь имеют связи шкал Профессий с Общими профессиональными темами, а также взаимосвязи между самими этими темами, показанные на рис. 14–2. Разработанная Холландом (Holland, 1966, 1985/1992) модель *R-I-A-S-E-C* имела огромное эвристическое значение в изучении профессиональных интересов, а инвентари Стронга, в свою очередь, были неотъемлемой частью этой исследовательской традиции, разумеется, вместе с другими средствами измерения интересов, использующими ту же схему. Результаты исследований с инвентарями Стронга, равно как и с применением других опросников, в целом довольно хорошо согласуются с предсказаниями, сделанными на основе модели Холланда.

Структура и организация *SII* позволяла модифицировать и расширять этот опросник в ходе его непрекращающихся пересмотров. Инвентарь Стронга не только первопроходческий инструмент в области измерения интересов, но и самый широко используемый психологами-консультантами в США (Watkins, Campbell, & Nieberding, 1994). Мало найдется тестов, которые использовались бы так долго и так широко. Тем не менее начавшийся с 1960-х гг. период отмечен быстрым приростом новых инструментов в этой области. Отчасти, этот процесс отражает увеличивающееся внимание к

изучению возможной карьеры и осознание того, что интересы играют ключевую роль в таком изучении.

Рассматриваемые в главе 17 комплексные программы изучения карьеры, как правило, учитывают меры профессиональных интересов, которые используются в сочетании с показателями тестов множественных способностей и информацией профессиографического характера. Например, данные об интересах, установленные с помощью Инвентаря профессиональных интересов (*Career Interest Inventory* [*CII*]), могут использоваться в связке с Дифференциальными тестами способностей (*Differential Aptitude Tests* [*DAT*]), вместе с которыми этот опросник был стандартизован (Psychological Corporation, 1991a). Аналогично, Программа обследования профессиональных способностей и интересов-2 (*Occupational Aptitude Survey and Interest Schedule-2* [*OASIS-2*] — Parker, 1991a, 1991b) создавалась для оценки старшеклассников по показателям 6 специальных способностей и 12 видов интересов в программах профориентации и подготовки к трудовой жизни. Еще одним примером инструмента, используемого в области планирования карьеры, может служить Пересмотренная система принятия карьерных решений Харрингтона—О'Ши (*Harrington—O'Shea Career Decision-Making System — Revised* [*CDM-R*] — *Harrington, & O'Shea, 1993*), *в которой предприняты* попытки интегрировать данные стандартизованных самоотчетов по интересам, ценностям и способностям с профессиографической информацией. Обозрение интересов и умений Кэмпбелла (*Campbell Interest and Skill Survey* [*CISS*] — Campbell, Hyne, & Nilsen, 1992) также измеряет интересы и умения на основе данных самоотчета и по своей организации напоминает *SII*, с которым Дэвид Кэмпбелл, автор *CISS*, работал какое-то время. Добавление данных по умениям позволяет проводить сравнения паттернов высоких и низких показателей по шкалам интересов и умений, что, в свою очередь, расширяет информационную базу поиска и выбора подходящей карьеры, обеспечиваемую *CISS*. Критические обзоры по *CII*, *OASIS-2*, *CDM-R* и *CISS* можно найти в работе Kapes et al. (1994).

Инвентари интересов: общий обзор и некоторые отличительные признаки

Среди множества ныне доступных инвентарей интересов четыре выбраны для особого упоминания по той причине, что каждый из них иллюстрирует какую-то примечательную особенность его теоретической ориентации, методологии или популяции, для которой он предназначен. Не следует ожидать от нас подробной характеристики или оценки этих инструментов, тем более что все они — либо в современных, либо в прежних своих версиях, — не так давно рецензировались в *Ежегодниках психических измерений (MMY)*.

Обозрение профессиональных интересов Джексона (Jackson Vocational Interest Survey [*JVIS*]). *JVIS* (Jackson, 1977) сделан предметом особого внимания, во-первых, потому что он представляет собой пример инструмента, созданного с помощью изощренных методов конструирования тестов, и, во-вторых, потому что подход к его созданию резко контрастирует с подходом, примененном при разработке *SII*. Разрыв между датами первой публикации опросников Стронга и Джексона составляет 50 лет.

Фокус инвентаря Стронга — конкретные профессии, причем этот фокус характерен как для отбора формулировок пунктов опросника, так и для нормативных интерпретаций; в опроснике Джексона используются широкие области интересов как при формулировке пунктов, так и при разработке системы оценивания результатов. Шкалы Профессий *SII* являют собой чистый образец привязки к эмпирическому критерию и валидизации относительно внешнего критерия; *JVIS* может служить классическим примером применения методов валидизации конструктов на всех этапах его разработки. В инвентаре Стронга подавляющее большинство пунктов предполагает независимые ответы респондента в форме «Нравится», «Безразлично» или «Не нравится»; в опроснике Джексона все пункты относятся к типу заданий с вынужденным выбором.

Так же как при разработке *PRF* и *JPI*, описанных в главе 13, первым шагом в создании *JVIS* было определение конструктов или измерений (*dimensions*), подлежащих оценке. Эти измерения, выбиравшиеся на основе опубликованных исследований по психологии труда и на основе факторно-аналитической и логической классификаций профессиональных интересов (в форме предполагаемых пунктов опросника), были двух видов. Одно измерение определялось через рабочие роли, а другое — через типы работы. *Рабочие роли (work roles)* относятся к тому, что человек делает на своем рабочем месте. Некоторые из этих ролей тесно связаны с конкретной профессией или ее типом, таким как инженерное дело, юриспруденция или начальное образование. Другие роли, такие как управление человеческими отношениями или профессиональное консультирование, являются сквозными, характерными для многих профессий. *Типы работы (work styles)* относятся не к связанной с конкретной должностью деятельности, а к предпочтениям в отношении рабочей среды или обстановки, предполагающей поведение определенного рода. Как правило, измерения (*dimensions*) типов работы либо прямо, либо косвенно связаны с ценностями человека. Примеры типов работы включают планомерную (без авралов), самостоятельную (не связанную с прямым участием других) и руководящую.

Разработка *JVIS* осуществлялась в несколько этапов, включая последовательные проверки и статистический анализ пунктов опросника, которые заготавливались в соответствии с детальной спецификацией для каждой рабочей роли и каждого типа работы. Отправляясь от исходного банка формулировок более 3000 пунктов, первоначально предъявлявшихся поодиночке для ответов в формате «нравится/не нравится», разработчики проводили факторный анализ подмножеств пунктов, подготовленных для каждой шкалы. При наличии систематической ошибки в ответах, обнаруживавшей себя в большом общем факторе, она устранялась статистически до того, как переходили к дальнейшему анализу пунктов. В итоге отбирались пункты, которые обнаруживали высокие корреляции с суммарными оценками фактора по своей собственной шкале и низкие корреляции с другими шкалами. Пары пунктов, представляющих различные рабочие роли или типы работы, частота одобрения которых оказалась примерно одинаковой при их одиночном предъявлении, переводились в формат задания с вынужденным выбором.[1]

Окончательная форма *JVIS* содержит 34 шкалы основных интересов, охватывающие 26 рабочих ролей и 8 типов работы. Опросник сконструирован таким образом,

[1] Однако в противоположность другим инструментам, в которых используются задания с вынужденным выбором, пары пунктов *JVIS* скомпонованы таким образом, что результирующие показатели *не* являются ипсативными по природе.

чтобы быть в равной мере применимым к представителям обоих полов, хотя имеются также отдельные процентильные нормы для мужской и женской подгрупп. Нормы были получены на больших выборках старшеклассников и студентов колледжей США и Канады. Высокий показатель по любой из 34 шкал основных интересов *JVIS* указывает на наличие интереса к тому, что люди делают в определенной сфере труда, а также к тому образу действий, который предполагается типичной для данной работы средой.

JVIS предусматривает возможность быстрой и удобной ручной обработки результатов по всем 34 шкалам. Однако имеющиеся в наличии варианты компьютерной обработки результатов *JVIS* используют более свежие нормы и обеспечивают несколько дополнительных видов анализа показателей как в форме короткого отчета, так и в недавно переработанной форме распространенного описательного заключения. Последнее содержит индивидуализированное описание и интерпретацию результатов респондента, а также большой объем сведений, способных оказать помощь при изучении возможной карьеры. Например, машинно-генерируемые отчеты включают показатели, выводимые на основе факторного анализа 34 шкал основных интересов. Эти показатели, сделанные по образцу шести тем Холланда, охватывают следующие 10 Общих профессиональных тем: Экспрессивную (*Expressive*), Логическую (*Logical*), Исследовательскую (*Inquiring*), Практическую (*Practical*), Ассертивную (*Assertive*), Социальную (*Socialized*), Помогающую (*Helping*), Конвенциальную (*Conventional*), Предпринимательскую (*Enterprising*) и Коммуникативную (*Communicative*). Кроме того, приводятся еще несколько показателей: мера Удовлетворенности образованием (*Academic Satisfaction*), индекс Согласованности ответов (*Response Consistency*), индекс Редкости (*Infrequency*) и количество Необрабатываемых ответов (*Unscorable Responses*).

Другие доступные виды анализа включают сравнения профиля респондента с модальными профилями, полученными в 17 кластерах академических областей специализации студентами колледжей и в 32 профессиональных кластерах работающими по найму лицами. Профессиональные профили были получены в результате совместного проведения *JVIS* и инвентарей Стронга, что позволило, применив факторный анализ, состыковать эти два инструмента. Установленная между ними связь открывает возможности использования при интерпретации показателей *JVIS* обширной базы данных, полученных с помощью инвентарей Стронга в отношении широкого круга профессий.[1]

В некоторых критических обзорах высказывались предположения, что формулировки пунктов *JVIS* могут быть слишком сложными для большинства учеников средней школы (D. T. Brown, 1989; J. W. Shepard, 1989). Недавно эти формулировки в буклете *JVIS* для респондентов были подвергнуты некоторой редакции, однако она имела целью обновление используемой терминологии, а не ее упрощение. Вскоре предстоит также пересмотр руководства по *JVIS* с целью включения норм, собранных в середине 1990-х гг. В настоящее время у издателей *JVIS* можно приобрести справочник по применению этого инструмента в профконсультировании (Verhoeve, 1993) и справочник по профессиям (Jackson, 1995).

[1] Полное описание процедур состыковки *JVIS* и инвентаря Стронга см. в Jackson (1977, chap. 4), Jackson, & Williams (1975).

Обозрение профессиональных интересов Кьюдера и его предшественники. Опросники для оценки интересов, разработанные Фредериком Кьюдером, вошли в употребление почти так же давно, как и комплект Стронга. Самым первым был Протокол профессиональных предпочтений Кьюдера (*Kuder Preference Record–Vocational* [*KPR-V*]). Примененный в нем подход к измерению интересов отличался от подхода Стронга в двух существенных отношениях. Во-первых, Кьюдер использовал пункты-триады с вынужденным выбором, отвечая на которые респонденты должны были указать, какое из трех занятий им нравилось больше всего, а какое — меньше всего. Во-вторых, показатели получались не для конкретных профессий, а для 10 широких областей интересов, именно: Работа на открытом воздухе (*Outdoor*), Работа с машинами и механизмами (*Mechanical*), Вычисления и расчеты (*Computational*), Научная работа (*Scientific*), Убеждение и побуждение других (*Persuasive*), Изобразительное искусство (*Artistic*), Литература (*Literacy*), Музыка (*Music*), Сфера социальных услуг (*Social service*) и Канцелярская работа (*Clerical*). Задания сначала формулировались и предварительно группировались исходя из содержательной валидности. Окончательный отбор заданий осуществлялся на основе внутренней согласованности и низких корреляций с другими шкалами. Позднее, в результате переработки и расширения возрастного диапазона KPR-V вниз появилось Обозрение общих интересов Кьюдера (*Kuder General Interest Survey* [*KGIS*]). Предназначенная для 6–12-х классов, эта форма использует более простую грамматику и лексику, требуя умения читать на уровне 6-го класса школы. Рецензии на *KGIS* (M. Pope, 1995; D. Thompson, 1995) можно найти в 12-м выпуске Ежегодника психических измерений (12^{th} *MMY*).

Еще более поздняя версия — Обозрение профессиональных интересов Кьюдера (*Kuder Occupational Interest Survey* [*KOIS*]) — дает показатели относительно конкретных профессиональных групп, как и инвентарь Стронга (Kuder, 1966; Kuder, & Diamond, 1979; Kuder, & Zytowski, 1991). Но, в отличие от *SII*, *KOIS* не предполагает использование ГЭВ. Вместо этого показатель респондента по каждой шкале Профессий выражается в виде корреляции между его паттерном интересов и паттерном интересов данной профессиональной группы.[1] Этот опросник может обрабатываться на сайте или в специальных пунктах издателей методом оптического сканирования бланков для ответов. Его также можно проводить и обрабатывать результаты непосредственно на компьютере. В настоящее время *KOIS* предусматривает показатели для 109 конкретных профессиональных групп и 40 областей специализации в колледжах. Некоторые шкал *KOIS* были разработаны только для мужчин, некоторые — только для женщих, а часть — для представителей обоих полов. Однако респондентам, независимо от пола, сообщаются показатели по всем шкалам. Охватываемые этим опросником профессии широко различаются по своему уровню, от косметолога и шофера грузовика до химика и адвоката.

Благодаря тщательному статистическому анализу показателей 3000 человек (по 100 в каждой из 30 основных групп представителей типичных профессий и областей специализации в колледже, охватываемых данным инвентарем) Кьюдеру удалось доказать, что лучшая дифференциация профессиональных групп достигается именно

---

[1] В качестве оценки такой корреляции используется коэффициент лямбда, разработанный Клемансом (Clemans, 1958). По существу, это точечно-бисериальный коэффициент корреляции, приспособленный к неоднородности разных критериальных групп. Дихотомическая переменная образована реакциями респондента (выбрал/не выбрал), отмеченными на бланке для ответов. Непрерывной переменной является доля лиц в критериальной группе, выбирающих каждый ответ.

при системе подсчета показателей, принятой в *KOIS*, чем при применении шкал Профессий, построенных с использованием ГЭВ. Исследования шкал других профессий продолжаются (Zytowski, 1992).

В настоящее время *KOIS* дает *и* показатели интереса к отдельным профессиям, *и* показатели интереса к 10 широким, однородным группам профессий, называемые Оценками профессиональных интересов (*Vocational Interest Estimates* или, сокращенно, *VIE*). Получаемые на основе коротких шкал VIE являются процентильными показателями, эквивалентными показателям 10 областей интересов *KPR-V*. Их также можно преобразовать в коды тем холландовской модели *R-I-A-S-E-C* путем прямого соотнесения шкал в одних случаях или путем усреднения процентилей по двум или трем шкалам Кьюдера в других (например, усреднить процентили по кьюдеровским шкалам интереса к областям изобразительного искусства, литературы и музыки для преобразования в холландовскую художественную шкалу). Рецензенты всегда хвалили технические характеристики *KOIS*, указывая при этом на нехватку данных о прогностической валидности этого инструмента. Кроме того, критике подвергалось невнимание разработчиков к воздействиям формата заданий с вынужденным выбором на показатели *KOIS* (см. Herr, 1989; Tenopyr, 1989).

Вопросник для оценки карьеры — Профессионально-техническая версия (*Career Assessment Inventory — The Vocational Version* [*CAI-VV*]).[1] Впервые опубликованный в 1975 г., *CAI* (Johansson, 1986) сделан по образцу инвентаря Стронга. Однако в противоположность большинству инвентарей интересов он разрабатывался специально для лиц, стремящихся найти род занятий, который *не* требовал бы четырехгодичного обучения в колледже или углубленной профессиональной подготовки. *CAI* сосредоточен на профессиях, требующих квалифицированного труда в сферах торговли, обслуживания, промышленного производства и эксплуатации оборудования. Примерами ныне имеющихся шкал *CAI* могут служить шкалы профессий авиамеханика, гигиениста в области стоматологии, работника кафетерия, программиста и аттестованной сиделки. 305 пунктов вопросника сгруппированы в три содержательные категории: Занятия (*Activities*), Школьные предметы (*School Subjects*) и Профессии (*Occupations*). Каждый пункт предоставляет на выбор пять вариантов ответа, от «очень нравится» до «очень не нравится». Поскольку требования *CAI* к навыкам чтения не превышают уровня 6-го класса школы, его можно также использовать в работе с плохо умеющими читать взрослыми. Подобно инвентарю Стронга, *CAI* дает показатели по трем главным типам шкал, включая 6 шкал Общих профессиональных тем Холланда, 22 шкалы однородных областей Основных интересов и 91 шкалу Профессий. Кроме того, он дает набор Административных индексов (*Administrative Indices*) и включает четыре шкалы, не имеющих отношения к профессиональным интересам (*Nonoccupational scales*).

Несмотря на то что примененные при разработке *CAI* методы имели большое сходство с процедурами, использованными при создании инвентарей Стронга, сбор всех данных по этому вопроснику и их статистический анализ проводился совершенно

---

[1] В настоящее время доступны две версии *CAI*: Профессионально-техническая версия (*VV*) и Расширенная версия (*The Enhanced Version* или, сокращенно, *EV*). В этом разделе речь идет только о *VV*. *EV*, несмотря на свое почти полное структурное сходство с *VV*, является абсолютно самостоятельным инструментом (Johansson, 1986), применимым к большему числу и более широкому диапазону профессий, включая и те, что требуют продолжения образования после окончания средней школы.

независимо. Поэтому, за исключением шкал Общих профессиональных тем, все остальные конкретные шкалы, разработанные для *CAI*, являются сугубо специфическими для данного вопросника. Что касается рецензий на *CAI-VV*, см. Kehoe (1992) и Vacc (1992).

Самоанализ профессиональных склонностей (*Self-Directed Search* [*SDS*]). Иллюстрацией другого подхода к оценке профессиональных интересов служит *SDS*. Этот инструмент был разработан Холландом, чья шестиугольная модель Общих профессиональных тем (обсуждавшаяся ранее в этой главе) привлекла к себе широкое внимание и получила применение в нескольких современных опросниках (Holland, 1985/1992; Holland, Fritzsche, & Powell, 1994; Holland, & Gottfredson, 1976; Holland, Powell, & Fritzsche, 1994).

SDS, как следует из его названия, создавался как применяемый к себе, самостоятельно обрабатываемый и интерпретируемый профориентационный тест. Хотя эта процедура организована вокруг интересов, она также предусматривает самооценку способностей и косвенные сведения о компетентности. Проводящий *SDS* заполняет Самооценочный буклет (*Self-Assessment Booklet*), подсчитывает ответы и вычисляет шесть суммарных показателей, соответствующих профессиональным темам модели Холланда (Реалистической, Исследовательской, Художественной, Социальной, Предпринимательской и Конвенциальной). Три самых высоких суммарных показателя используют для нахождения трехбуквенного кода.[1] Входящий в комплект *SDS* Определитель профессий (*Occupational Finder*) позволяет отыскать среди 1335 профессий те, чьи коды имеют сходство с итоговым кодом респондента. Включенные в этот определитель профессии отбирались таким образом, чтобы представить почти всех трудящихся в США; преобразовав код, пользователь может расширить границы поиска, используя *Словарь названий профессий (Dictionary of Occupational Titles)*. При желании предоставляются дополнительные инструкции, методики и источники информации, облегчающие пользователю принятие решения о выборе профессии.

SDS нашел широкое применения в самых разнообразных контекстах и породил обширные исследования, проводимые как его создателем, так и независимыми специалистами. За годы, прошедшие с момента его публикации, он подвергался неоднократным пересмотрам в целях упрощения процедуры и уменьшения влияния пола респондента на принятие решений о выборе карьеры. Практическая привлекательность *SDS* заключена в его краткости и простоте, возможности самоприменения и способности существенно расширить поле выбора карьеры его пользователем. В добавление к основной форме вопросника (*Form R*, от англ. *Regular*), сейчас пользователям доступны три других его версии: 1) Упрощенная форма (*Form E*, от англ. *Easy*), разработанная для лиц с ограниченными навыками чтения; 2) Форма для планирования карьеры (*Form CP*, от англ. *Career Planning*), предназначенная для взрослых, оказавшихся пере выбором в связи с необходимостью поменять работу, и 3) Путеводитель по профессиям (*Career Explorer*), предназначенный для учащихся 6–10-х классов американской школы.

Если говорить о психометрических качествах *SDS*, показатели надежности которых для суммарных показателей обычно оказываются удовлетворительными. Вали-

---

[1] Хотя *SDS* предполагает самостоятельную обработку ответов, руководство к нему рекомендует все же некоторые формы надзора за процессом обработки и проверки получаемых показателей. Исследование 107 случайно выбранных респондентов разного возраста, проводивших *SDS* в ныне существующей редакции, показало, что 7,5 % получили коды, содержащие ошибочные буквы или их транспозиции (Holland, Powell, & Fritzsche, 1994, p. 16).

дизация конструктов основных шести профессиональных тем основывается, главным образом, на исследовании, которое и привело к их формулированию, а также на проведенных позднее конфирматорных факторно-аналитических исследованиях, в основном подтвердивших правомерность выделения шести этих тем (см., например, Oosterveld, 1994). Текущая валидность и предсказуемостная эффективность *SDS* колеблется в зависимости от состава выборок, рассматриваемого с точки зрения возраста, пола, уровня образования и распределения типов (Holland, Fritzsche, & Powell, 1994). Критика *SDS* касается некоторых из используемых в нем процедур подсчета баллов и интерпретации показателей (M. H. Daniels, 1989; Manuele-Adkins, 1989). Тем не менее рецензенты сходятся в том, что этот инструмент дает в наше распоряжение простой, недорогой и относительно точный способ изучения профессиональных склонностей. Издатели предоставляют также библиографию по *SDS* и ряд информационных материалов, используемых в сочетании с этим вопросником.

Холланд (Holland, 1966, 1985/1992) открыто относит себя к числу тех, кто рассматривает профессиональные предпочтения как выбор образа жизни — выбор, который отражает представления человека о себе (Я-концепцию) и своеобразие его личности. Каждая из профессиональных тем Холланда соответствует «типу» или кластеру свойств личности. Склонного к определенной профессиональной деятельности человека можно охарактеризовать через указание одного или большего числа преобладающих типов. Кроме того, темы Холланда соответствуют идеализированным средам, через которые можно описывать обстановку, царящую на рабочем месте представителей разных профессий. Эти среды включают в себя не только физические характеристики и требования, предъявляемые к человеку конкретной работой, но и категории лиц, с которыми он работает (сотрудников, начальников, покупателей, клиентов, учащихся). Согласно Холланду, каждый человек стремится попасть в среду, конгруэнтную его типу личности, и такая конгруэнтность повышает удовлетворенность работой, ее стабильность и успешность.

Рассматриваемый в другой перспективе, общий подход Холланда к оценке профессиональных интересов согласуется с некоторыми важными выводами в психологии выбора профессии. Д. Супер (Super, 1953, 1957, 1990) неоднократно утверждал, что выбор профессии — это реализация Я-концепции, т. е. представлений человека о самом себе. Направление исследований, нацеленных на выявление особенностей личности представителей разных профессиональных групп, имеет давнюю историю (см., например, Borgen, 1986; Costa, McCrae, & Holland, 1984; Osipow, 1973, chap. 6; Pietrofesa, & Splete, 1975, chap. 4; Super, & Bohn, 1970, chap. 5). Выбор профессии часто отражает базисные эмоциональные потребности индивидуума. А профессиональная адаптация является главной составной частью общего приспособления к жизни (Tait, Padgett, & Baldwin, 1989). Оценка профессиональных интересов — и более конкретно, выявление тех профессиональных групп, чьи интересы и аттитюды наиболее близки индивидууму, — становится таким образом фокальной точкой в понимании различных личностей.[1]

---

[1] Wallet, Lykken и Tellegen (1995) сообщают о некоторых результатах, касающихся взаимосвязей между профессиональными интересами, увлечениями в свободное от работы время и чертами личности, обнаруженных в большой выборке участников Миннесотского регистра близнецов (*Minnesota Twin Registry*). Их исследования говорят о том, что, хотя черты личности явно опосредуют соответствие «человек—профессия», три множества соответствующих характеристик могут концептуализироваться как особые области.

Некоторые важные тенденции

Разработка и использование опросников. К самым ярким событиям в области современных средств измерения интересов можно отнести слияние двух главных теоретических позиций в психологии профессий (*occupational psychology*), а также и различных подходов к конструированию опросников, и перекрестное использование банков эмпирических данных в целях интерпретации. Все больше и больше инструментов обеспечивают получение показателей *как* по однородным, общим шкалам интересов, *так* и по специфическим шкалам профессий. Шесть профессиональных тем модели Холланда вновь «всплыли на поверхность» в большом количестве недавно разработанных или пересмотренных инвентарей интересов.

С другой стороны, устанавливаются связующие звенья с имеющимися эмпирическими данными о профессиях, накопленными в отношении больших групп населения. Эта тенденция была проиллюстрирована на примере использования данных по инвентарю Стронга при интерпретации результатов, полученных с помощью Обозрения профессиональных интересов Джексона. Ее иллюстрацией служит также связывание многих современных инвентарей, таких как *SDS* Холланда и *SII*, с данными, предоставляемыми *Словарем названий профессий* (U. S. Department of Labor Employment and Training Administration, 1991). Все это — многообещающие события, которые повышают полезность любого отдельно взятого инструмента, и их последствия должны быть благотворными, разумеется, при условии, что при установлении таких связей применяются надежные психометрические процедуры и принимаются необходимые меры для предотвращения возможной сверхгенерализации в интерпретации показателей.

Другая отличительная особенность вновь создаваемых или перерабатываемых в последнее время инвентарей интересов — расширение уровней охватываемых ими профессий. Вначале инвентари интересов были нацелены на выбор карьеры, связанной с интеллектуальным трудом, и включали малое число профессий, которые не требовали образования на уровне колледжа или профессиональной школы. Несмотря на отдельные попытки расширить в инвентарях интересов диапазон профессий в сторону менее квалифицированного труда (например, Clark, 1961), эти инструменты не нашли в то время широкого применения. В противоположность этому, самая свежая версия инвентаря Стронга предлагает значительное число шкал профессий в разных сферах труда, не требующих диплома об окончании колледжа. Более того, некоторые из последних разработок, такие как Вопросник для оценки карьеры — Профессионально-техническая версия (Johansson, 1984) и Инвентарь направлений профессиональной деятельности (*Career Directions Inventory* — Jackson, 1986), предназначены прежде всего для лиц, чье образование не выходит за рамки средней школы или профессионально-технического училища. Скорее всего, такие изменения отражают, по крайней мере отчасти, растущее осознание важности правильного выбора карьеры на всех профессиональных уровнях и вместе с тем решающей роли интересов в достижении успеха и личной удовлетворенности работой независимо от ее характера.

Еще одна заметная тенденция имеет последствия, выходящие за пределы сферы измерения интересов и распространяющиеся на другие виды тестирования. В своей смелой попытке предугадать будущее инвентарей интересов Дж. Л. Холланд (Holland, 1986) указал на растущее понимание таких опросников, как методик вмешательства. Какие *воздействия* оказывает инвентарь интересов на тестируемого? Например, у од-

ного человека он может укрепить и усилить существующие профессиональные стремления. Другого он может побудить к всестороннему изучению мира профессий и привлечь внимание к не рассматриваемым до сих пор вариантам карьеры. Ну а третьему просто помочь лучше разобраться в себе. Такое разнообразие возможных воздействий может отражаться не только в индивидуализированных интерпретациях показателей, но также в проведении и даже конструировании инвентарей интересов. Вполне вероятно, что здесь находится еще одна сфера приложения компьютеризованного адаптивного тестирования.

В 1990-х гг. психология профессионального самоопределения и трудовой жизни обогатилась за счет применения парадигм из области когнитивной психологии (см., например, Peterson, Sampson, & Reardon, 1991). Взгляд на принятие связанных с профессией решений как на процесс решения задач (*problem-solving*), возобновляемый на разных этапах жизненного пути человека, подчеркнул потребность людей в знаниях и навыках обработки информации, необходимых для достижения оптимальных решений. Многие новые инструменты встроены в комплексные программы исследования карьеры, являющиеся логическим продолжением взглядов Д. Супера на профессиональный цикл человека и подхода Дж. Л. Холланда к выбору профессии. Становится все более очевидным, что область профессионального поведения (*vocational behavior*)[1] предлагает исключительно благоприятные возможности для вовлечения людей в активное и непосредственное пользование теоретическими и инструментальными достижениями психологии (Borgen, 1991).

Модели профессий. Большинство рассматриваемых в этой главе средств измерения профессиональных интересов либо построены на основе холландовской модели профессиональных тем (например, *CAI* и *SDS*), либо вобрали ее — в большей или меньшей степени — в свои процедуры (например, *SII* и *KOIS*). Фактически, на протяжении двух последних десятилетий теоретические постулаты Холланда играли ведущую роль в стимулировании исследований в области психологии профессий, причем не только в США, но и в других частях мира (см., например, Borgen, 1991; Lokan, & Taylor, 1986). Осуществляя всестороннюю проверку этой модели, Трейси и Раундс (Tracey, & Rounds, 1993) провели на широком множестве инструментов структурный метаанализ шкал *R-I-A-S-E-C*. Это исследование, охватившее 104 американские выборки, дало хорошее подтверждение модели Холланда. Однако аналогичный метаанализ, проведенный позднее теми же авторами с использованием иностранных и американских выборок этнических меньшинств, не показал столь же хорошего соответствия эмпирических данных проверяемой модели (Rounds, & Tracey, 1996). Другие сравнительные исследования разных национальных и этнических выборок дали смешанные результаты (Fouad, & Dancer, 1992; Hansen, 1987; Khan, Alvi, Shaukar, & Hussain, 1990; Swanson, 1992). Пожалуй, не стоило рассчитывать на то, что кросскультурные исследования модели *R-I-A-S-E-C* покажут ее одинаковую применимость, причем во всех своих аспектах и в разных культурах.

Так или иначе, простота шестиугольника *R-I-A-S-E-C* и явное господство этой модели в исследованиях последних 20 лет склонили многих специалистов к выводу, что, возможно, для этой области настало время выйти за пределы модели Холланда или,

[1] Имеется в виду поведение, связанное с реализацией задач, встающих пред человеком на всех этапах профессионального цикла (см., например, Super, 1957). — *Примеч. науч. ред.*

по крайней мере, повысить ее полезность путем введения дополнительных элементов.[1] Одни призывали к разработке новых теоретических структур для объяснения других аспектов профессий, таких как способности и предпочитаемые источники подкрепления, тогда как другим хотелось бы видеть исследованными дополнительные переменные и измерения (см., например, Dawis, 1992; Prediger, & Vansickle, 1992; R. H. Schwartz, 1992). Сам Холланд заявил, что исследования следует нацелить на изучение способов применения и интерпретации инвентарей интересов, а также оказываемого ими влияния на тестируемых, а не на бесконечное воспроизведение шестиугольника профессиональных тем в различных выборках (Holland, & Gottfredson, 1992).

Одна из самых амбициозных попыток модифицировать и расширить модель Холланда с помощью новых и весьма сложных методов анализа данных, включая многомерное шкалирование, была предпринята Трейси и Раундсом (Tracey, & Rounds, 1996). Их трехмерное представление профессиональных интересов использует типологию Холланда и еще два измерения (*dimensions*), которые, предположительно, лежат в ее основе,[2] как отправную точку для построения расширенной сферической модели, приспосабливаемой к переменному числу типов интересов, в зависимости от потребностей тестируемого, и вмещающей такое измерение, как *престиж (prestige)*. Недавний номер *Journal of Vocational Behavior* специально посвящен представлению этой новой модели и комментариям к ней разных исследователей. Реакции на работу Трейси и Раундса, в общем, положительные и свидетельствуют о том, что их модель, по-видимому, служит своего рода катализатором интеграции и дальнейших прорывов в области теории выбора профессии и измерений профессиональных интересов (см. Borgen, & Donnay, 1996; Gonzalez, 1996; G. D. Gottfredson, 1996; Hansen, 1996; Harmon, 1996; Prediger, 1996).

Исследования характера, количества и структуры базисных интересов имеют сходство с исследованиями, проводимыми с целью установления главных (первичных) факторов в области способностей и личности (см. главы 11 и 13).[3] Во всех трех случаях выявленные в результате анализа данных категории являются функцией специфических переменных и тех выборок, на которых собирались эти данные. Кроме того, это описательные, а не объяснительные категории, основная ценность которых заключается в их способности упрощать сбор и использование информации для оценки и предсказания поведения.

Опросы мнений и шкалы аттитюдов

Сущность инструментария. Аттитюд часто определяют как склонность одобрительно или неодобрительно реагировать на определенный класс стимулов, таких как национальная или этническая принадлежность, обычай или общественное установление. Очевидно, что определяемые таким образом аттитюды не могут наблюдаться непо-

[1] Специальный выпуск *Journal of Vocational Behavior* (April 1992), на который мы несколько раз ссылаемся в этом разделе, был полностью посвящен обсуждению теории Холланда.

[2] Имеются в виду два выделенных Предигером (Prediger, 1982) и с тех пор широко используемых биполярных измерения: «люди/вещи» и «факты/идеи».

[3] Широкомасштабная и поучительная историческая картина попыток идентифицировать и классифицировать измерения (*dimensions*) интересов, а также отобразить их взаимосвязи, представлена в работе Раундса (Rounds, 1995).

средственно, но должны логическим путем выводиться из внешнего поведения, как вербального, так и невербального. Пользуясь более объективной терминологией, можно сказать, что понятие аттитюда подразумевает *постоянство реакции (response consistency)* относительно некоторых категорий стимулов. На практике термин «аттитюд» наиболее часто связывался с социальными стимулами и с эмоционально окрашенными реакциями. Кроме того, он нередко связывается с ценностными суждениями.

Иногда мнение отличают от аттитюда, но предлагаемые отличительные признаки обычно не выдерживают проверки на логическую непротиворечивость. Более часто эти два термина употребляются как взаимозаменяемые, и точно так же они будут использоваться в этом разделе. Что касается методологии оценки, то опросы мнений традиционно отличаются от шкал установок. *Опросы мнений (opinion surveys)* обычно имеют дело с ответами на конкретные вопросы, которые не обязательно связаны между собой. Ответы на такие вопросы не объединяются в суммарном показателе, а анализируются по отдельности. Опросный лист для изучения мнений наемных работников, например, может содержать вопросы о графике работы, ставках, дополнительных льготах, столовой и отношениях с начальством; каждый из этих пунктов включен в опрос вследствие его собственной значимости для улучшения взаимоотношений между членами производственного коллектива. Ответы на каждый вопрос сводятся в отдельные таблицы с тем, чтобы выявить источники удовлетворенности и неудовлетворенности работников.[1]

С другой стороны, *шкалы аттитюдов (attitude scales)* обычно дают суммарный показатель, указывающий направление и силу аттитюда конкретного человека в отношении промышленной компании, группы людей, политики или других стимульных категорий. При конструировании шкалы для оценки аттитюда различные вопросы придумываются для измерения единичного аттитюда или одномерной переменной, и для достижения этой цели обычно используют ряд объективных процедур. Шкала аттитюда наемного работника, например, дает единственный показатель, отражающий степень удовлетворенности конкретного лица работой или его общее отношение к компании.

Основные типы шкал аттитюдов. Во всех шкалах аттитюдов респонденты отмечают свое согласие или несогласие с серией утверждений об объекте аттитюда. Для достижения одномерности или однородности пунктов шкалы, равенства расстояний между единицами шкалы и сопоставимости показателей при переходе от одной шкалы к другой были разработаны специальные процедуры. Возникающим при конструировании шкал аттитюдов техническим проблемам уделялось широкое внимание, и потому эта область методологии измерений отмечена значительными теоретическими и статистическими разработками. В задачи этого учебника не входит обсуждение специализированных методик шкалирования, которые составляют сейчас развивающуюся область статистических методов (Jones, & Koehly, 1993; D. J. Mueller, 1986; Ostrom, Bond, Krosnick, & Sedikides, 1994; Procter, 1993; Reckase, 1990; Young, 1984). Однако мы можем кратко рассмотреть три основных подхода к конструированию шкал аттитюдов, чаще других встречающиеся в литературе по психологическому тес-

[1] Многотомная серия под редакцией Финка (Fink, 1995) служит исчерпывающим руководством по методам опроса, освещающим конкретные вопросы планирования, подготовки, проведения и анализа полученных данных.

тированию. Эти подходы представлены такими типами шкал, как шкалы Терстоуна, Гуттмана и Лайкерта.[1]

Тёрстоун приспособил психофизические методы для квантификации данных суждений, и это стало важной вехой в развитии конструирования шкал аттитюдов (Thurstone, 1959; Thurstone, & Chave, 1929). С помощью этих методов Тёрстоун и его сотрудники подготовили около 20 шкал для измерения аттитюдов в отношении войны, смертной казни, церкви, патриотизма, цензуры и многим другим общественным институтам, установленным порядкам, спорным вопросам и национальным (или этническим) группам. Разработка *шкалы тёрстоуновского типа (Thurstone-type scale)* начинается с собирания множества утверждений, выражающих широкое разнообразие аттитюдов в отношении рассматриваемого объекта. Затем большую группу экспертов просят индивидуально разложить эти утверждения по стопкам (обычно 11) по степени благосклонности. Отметим, что эксперты не выражают (или, по крайней мере, не должны выражать) свои собственные аттитюды; они только классифицируют утверждения. Медианное положение, определенное для каждого утверждения экспертами, и является шкальной оценкой данного утверждения. Изменчивость утверждений принимается за показатель ее неопределенности, поскольку разные эксперты относят одно и то же утверждение к разным категориям. В качестве пунктов шкалы выбираются такие утверждения, которые обнаруживают минимальную изменчивость и обеспечивают широкий разброс шкальных оценок, имеющих приближенно равноинтервальное распределение на 11-пунктной шкале. В окончательном варианте шкалы аттитюда эти утверждения предъявляются в случайном порядке, без указания их шкальных оценок. Показателем респондента служит медианная шкальная оценка всех одобренных им утверждений.

Шкала гуттмановского типа (Guttman-type scale) первоначально разрабатывалась как средство для определения того, является ли множество словесных формулировок аттитюда одномерным (L. Guttman, 1944, 1947). В понимании Гуттмана совершенная шкала существует в том случае, если респондент, соглашающийся с предельно ясной формулировкой какого-то конкретного аттитюда, соглашается и с более мягкими формулировками этого аттитюда. Другими словами, пункты такой шкалы аттитюда можно упорядочить на континууме его силы или трудности его принятия. Положение каждого человека на данной шкале полностью определялось бы его ответами. Если нам известна самая крайняя из принятых индивидуумом формулировок аттитюда, то в этом случае мы должны суметь воспроизвести все его ответы. На практике, однако, не удается достичь полной воспроизводимости из-за ошибок измерения в каждом ответе, можно лишь приблизиться к ней в известных пределах. Важнейшей процедурой при разработке шкалы Гуттмана является определение множества пунктов, которые образуют упорядоченную последовательность с точки зрения их одобрения респондентами. Не отвечающие этому требованию пункты отбрасываются. Показатель респондента по шкале Гуттмана выводится на основе исследования паттернов одобренных им пунктов. Здесь можно напомнить, что понятие ординальности, или равно-

[1] Доступное изложение технических подробностей конструирования этих и других типов шкал аттитюдов, без знания которых знакомство с данным видом психологических измерительных инструментов останется весьма поверхностным, отечественный читатель может найти в соответствующих разделах и главах двух (полезных во многих отношениях) книг: Математические методы в современной буржуазной социологии. — М.: Прогресс, 1966; Процесс социального исследования. — М.: Прогресс, 1975. — *Примеч. науч. ред.*

мерной прогрессии исполнения, лежит в основе шкал Пиаже, которые обсуждались в главах 3 и 9.

Исходя из того, что конструирование шкалы Тёрстоуна требует применения довольно сложных процедур, а параметры шкалы Гуттмана труднодостижимы на практике, Лайкерт (Likert, 1932) разработал тип шкалы, которая гораздо легче конструируется и при этом обеспечивает столь же удовлетворительную надежность измерений. *Шкала лайкертовского типа (Likert-type scale)* начинается с серии утверждений, каждое из которых выражает аттитюд, являющийся либо откровенно одобрительным, либо откровенно неодобрительным. Пункты шкалы отбираются на основе ответов тех лиц, которым они предъявляются в процессе ее конструирования. Ведущим принципом отбора пунктов выступает их внутренняя согласованность, хотя при возможности используются и внешние критерии. Шкалы Лайкерта требуют классификационного ответа на каждое утверждение. Обычно такой ответ выражается одной из следующих пяти категорий: полностью согласен (*SA*), согласен (*A*), не уверен (*U*), не согласен (*D*), полностью не согласен (*SD*). Для подсчета показателей шкалы вариантам ответов приписываются условные баллы 5, 4, 3, 2 или 1, от благоприятного до неблагоприятного полюса. Например, ответ «полностью согласен» в отношении благожелательного утверждения получил бы показатель 5, так же как и ответ «полностью не согласен» в отношении неблагожелательного утверждения. Суммирование этих условных баллов по пунктам шкалы дает общий показатель аттитюда индивидуума, который следует интерпретировать исходя из эмпирически установленных норм.

Большинство шкал аттитюдов было разработано для использования в конкретных исследовательских проектах. Одни из них предназначались для выяснения аттитюдов и морального духа наемных работников. Другие применялись для оценки результата образовательных программ и разного рода тренингов. Шкалы аттитюдов могут быть полезны при оценивании различных воспитательных процедур, предназначенных для изменения специфических аттитюдов. Или их можно использовать при измерении изменений аттитюдов учащихся в отношении литературы, искусства, различных этнических и культурных групп или социальных и экономических проблем, происходящих под воздействием определенной образовательной программы. Пожалуй, самое широкое применение измерение аттитюдов находит в исследованиях по социальной психологии. Практически любой учебник по социальной психологии содержит разделы по аттитюдам и их измерению. Среди множества проблем, исследуемых посредством измерения аттитюдов, можно упомянуть групповые различия в аттитюдах, роль аттитюдов в межгрупповых отношениях, биографические факторы развития аттитюдов, взаимосвязи аттитюдов (включая факторный анализ и другие методы многомерного анализа), тенденции изменения аттитюдов со временем, а также экспериментальная переделка аттитюдов на основе интерполированного опыта (*interpolated experiences*).[1] Сравнительно мало шкал аттитюдов издавалось для пользователей специализированными издательствами, хотя большинство из них полностью описаны в научно-исследовательской литературе. Обширная коллекция ранних шкал аттитюдов, конструировавшихся для самых разных целей, собрана в книге Шоу и Райта (M. E. Shaw & Wright, 1967). Сведения о более современных средствах измерения

[1] Блестящий обзор широкого направления социальной психологии, занимающегося аттитюдами, можно найти в работе Eagly, & Chaiken (1993). Более короткий обзор литературы по аттитюдам и их изменению см. у Olson, & Zanna (1993).

некоторых аттитюдов и ценностей, таких как отчуждение и аномия, самоуважение и локус контроля, можно найти в книге *Measures of Personality and Social Psychological Attitudes* (Robinson, Shaver, & Wrightsman (Ed.), 1991).

Замечание по поводу связанных с полом переменных и средств их измерения. Предыдущие издания этой книги включали разделы, освещающие с большими или меньшими подробностями средства оценки половых ролей и таких связанных с ними концептов, как маскулинность, фемининность и андрогиния, тогда как в этом издании их нет вообще. Хотя исследования в данной области, вместе с используемыми в них инструментами для оценки этих феноменов, продолжали быстро разрастаться (см., например, Lenney, 1991), многие исследователи сходятся в том, что эта область находится сейчас в состоянии концептуального беспорядка. В рамках более широкой перспективы индивидуальных различий в связанных с полом переменных Бец (Betz, 1995) подготовила великолепную обновленную сводку современного состояния знаний в отношении таких переменных. Она приходит к выводу, что отсутствие ясной теоретической основы и четких определений основных понятий стало причиной сдерживания прогресса в области изучения связанных с полом феноменов. К этому, безусловно, следовало бы добавить, что за последние два десятилетия произошли разительные изменения во взглядах на пол человека в культуре США и по всему миру. Во всяком случае, как отмечает Бец, такие широкие объяснительные понятия для связанных с полом различий, как половые роли или маскулинность/фемининность, почти совсем не получили эмпирической поддержки. Согласно Бец, для появления прогрессивных изменений в этой области исследований необходимо проделать большую работу, заключающуюся в тщательной концептуализации и определении конструктов, которые можно было бы вписать в обоснованную и связную теоретическую систему.

Локус контроля

Конструкт, описанный как «локус контроля», приобрел известность одновременно с публикацией монографии Джулиана Роттера (Rotter, 1966). В этой публикации Роттер представил шкалу, разработанную им для оценки генерализованных ожиданий внутреннего (*internal*)/внешнего (*external*) контроля подкрепления — так называемую шкалу И–Э (*I–E scale*). Этот инструмент был сконструирован в контексте теории социального научения. Объясняя его назначение, Роттер писал: «Эффект подкрепления, следующего за поведением,... не просто механический процесс, но зависит от того, воспринимает ли человек причинную связь между собственным поведением и вознаграждением» (1966, p. 1). Внутренний (интернальный) контроль указывает на восприятие события как обусловленного поведением или относительно постоянными характеристиками индивидуума. С другой стороны, внешний (экстернальный) контроль указывает на то, что следующее за действием индивидуума положительное или отрицательное подкрепление воспринимается им не как полностью зависящее от его действия, а как результат случая, рокового или счастливого стечения обстоятельств; или же оно может восприниматься как находящееся под контролем могущественных других и непредсказуемым из-за сложности сил, действующих в окружении данного человека.

Таблица 14–1

Два иллюстративных пункта шкалы И–Э

(*a*) В конечном счете, люди получают то признание, которого они заслуживают в этом мире. (*b*) К сожалению, заслуги человека часто остаются непризнанными, несмотря на все его труды и старания.
(*a*) Успех приходит в результате упорного труда, удача здесь не при чем — или почти не при чем. (*b*) Для получения хорошего места главное — оказаться в нужном месте в нужное время.

В инструкции говорится: «Этот опросник служит для выяснения того, каким образом некоторые важные события в нашем обществе влияют на различных людей... Пожалуйста, выберите из каждой пары одно *(и только одно)* утверждение, которое *вы считаете* более подходящим в том случае, когда речь идет о вас самих».

(Из Rotter, 1966, p. 11. Воспроизводится с разрешения)

Шкала И–Э представляет собой опросник, относящийся к типу самоотчетов с вынужденным выбором ответов. Два иллюстративных пункта из этого опросника приведены в табл. 14–1. Полный перечень пунктов опросника вместе со стандартными инструкциями по их применению можно найти в монографии Роттера (Rotter, 1966). Этот первоисточник содержит значительное количество сведений о шкале И–Э, включая процентильные нормы по нескольким сотням студентов мужского и женского пола из одного университета, вместе со средними и *SD* для дюжины других выборок, охватывающих большинство групп колледжа. Впоследствии были собраны данные по множеству других групп, полученные в рамках независимых исследовательских проектов. К настоящему времени накоплен также существенный объем данных по конструктной валидности. Первоначально факторный анализ показал, что большую часть дисперсии ответов можно объяснить одним общим фактором. Однако результаты последующих факторно-аналитических исследований указывают на то, что конструкт «локус контроля» можно подразделить на несколько различимых факторов, иллюстрируемых мнениями о трудном мире, несправедливом мире, непредсказуемом мире и политически безответственном мире (B. E. Collins, 1974). В нескольких более поздних исследованиях также была получена многофакторная структура конструкта «локус контроля».

В период с середины до конца 1970-х гг. было с большой степенью достоверности установлено, что ожидания контроля того типа, которые выявляются с помощью шкалы И–Э, могут играть важную роль в предсказании некоторых видов поведения. Однако работающие в этой области исследователи также признали, что для максимально возможного повышения точности предсказания необходимо приспособить средства измерения ожиданий контроля к специфическим популяциям и изучаемым областям поведения. С тех пор было разработано несколько различных шкал локуса контроля. Некоторые из них конструировались для использования в различных популяциях, включая дошкольников и младших школьников (см., например, Connell, 1985; Herzberger, Linney, Seidman, & Rappaport, 1979; Nowicki, & Duke, 1983). Другие нацелены на оценку мнений о контроле в таких специфических областях, как удовлетворенность браком или состояние психического здоровья (D. J. Hill, & Bale, 1980; P. C. Miller, Lefcourt, & Ware, 1983). Третьи охватывают каузальные представления, касающиеся

различных областей (например, достижения или аффилиации) или различных сфер контроля, таких как личная действенность, интерперсональный контроль и общественно-политический контроль (Lefcourt, von Baeyer, Ware, & Cox, 1979; Paulhus, 1983). Воспроизведение ряда имеющихся шкал, вместе с основными психометрическими данными, можно найти в работе Lefcourt (1991).

Количество и диапазон доступных инструментов для оценки конструкта «локус контроля» подтверждает его непрекращающееся применение. Более 5000 статей перечислено под этой рубрикой в базе данных психологической литературы PsychINFO за период с 1984 по 1995 г. Особенно много исследований посвящено роли ожиданий контроля в практике охраны и поддержания здоровья. Вдобавок ко всему, локус контроля — важный аспект мотивации вообще и тесно связан с другими основными областями исследования личности, включая, помимо прочего, теорию атрибуции, приобретенную беспомощность и самоэффективность (Skinner, 1995).

15 ПРОЕКТИВНЫЕ МЕТОДИКИ

В настоящее время мы располагаем достаточно большим набором разнообразных проективных методик. В этой главе будут рассмотрены основные виды таких методик, вместе с некоторыми широко известными их примерами. За исключением важных отличительных особенностей, присущих конкретным методикам, никакой критической оценки отдельным инструментам даваться не будет. Вместо этого в специальном разделе будет дана совокупная оценка проективных методов с акцентом на общих методологических проблемах. Проективные методики являют пример любопытного расхождения исследования и практики: оцениваемые как психометрические инструменты, они в подавляющем большинстве выглядят жалко, но такая оценка тем не менее никак не сказывается на их популярности в клиническом использовании (Bellak, 1992; Lubin, Larsen, & Matarazzo, 1984; Piotrowski, 1984; Piotrowski, Sherry, & Keller, 1985; Piotrowski, & Zalewski, 1993; Watkins, 1991). Природа и следствия этого несоответствия будут рассмотрены в последнем разделе главы.

Литература по проективным методикам обширна, насчитывая свыше 6000 источников по одному только инструменту. Для более полного представления об имеющихся проективных методиках читателю следует обратиться к работам Klopfer, & Taulbee (1976), Rabin (1981, 1986), Reynolds, & Kamphaus (1990b, chaps. 3–8). *Ежегодники психических измерений (MMY)* содержат критические разборы большинства современных инструментов. *Журнал оценки личности (Journal of Personality Assessment)*[1], хотя и принимает статьи по всем типам тестов, публикует богатый материал по различным аспектам использования проективных методик.

Природа проективных методик

Главную отличительную особенность проективных методик нужно искать в относительно *неструктурированной (unstructured)* задаче для испытуемого, т. е. задаче, допускающей почти неограниченное разнообразие возможных ответов. Для того что-

---

[1] Это издание несколько раз меняло название, с тех пор как в 1936 г. было учреждено Бруно Клопфером как информационный бюллетень *Rorschach Research Exchange*; с 1950 по 1963 г. это был *Journal of Projective Techniques (Журнал проективных методик)*.

бы фантазия индивидуума могла свободно разыграться, даются только краткие, общие инструкции. По этой же причине тестовые стимулы обычно расплывчаты или неоднозначны. Основная гипотеза создателей проективных методик состоит в следующем: то, каким образом конкретный человек воспринимает и интерпретирует тестовый материал или «структурирует» ситуацию, по-видимому, отражает фундаментальные аспекты функционирования его психики. Другими словами, предполагается, что тестовые материалы служат своего рода экраном, на который респонденты «проецируют» свои характерные мыслительные процессы, потребности, тревоги и конфликты.

В типичных случаях проективные инструменты представляют собой методики *замаскированного (disguised)* тестирования, поскольку обследуемый редко подозревает о типе психологической интерпретации, которая будет дана его ответам. Проективные методики характеризуются также *глобальным (global)* подходом к оценке личности. Внимание фокусируется на составной («композитной») картине личности в целом, а не на измерении отдельных черт. Наконец, приверженцы проективных методик обычно считают их особенно эффективными при выявлении *скрытых, латентных* или *неосознаваемых* сторон личности. Более того, утверждается, что чем менее структурирован тест, тем он более чувствителен к такому скрытому материалу. Это следует из предположения, что чем менее структурированы и однозначны стимулы, тем менее вероятно, что они вызовут у респондента защитные реакции.

Проективные методы были созданы в клинических условиях и оставались в основном инструментом клинициста. Некоторые из них развились из терапевтических методов (таких, как арт-терапия), применявшихся к психически больным. Что касается теоретической основы большинства проективных методик, то здесь явно заметно влияние традиционных и современных психоаналитических концепций. Предпринимались также разрозненные попытки заложить фундамент для проективных методик в стимульно-реактивной теории и в перцептуальных теориях личности[1] (см., например, Lindzey, 1961/1977). Следует, конечно, отметить, что *нет* необходимости оценивать конкретные методики в свете теоретического подхода к их созданию или их исторических корней. Методика может оказаться практически полезной или эмпирически ценной по другим причинам, нежели те, которые первоначально приводились для оправдания ее введения.

Методики чернильных пятен

Тесты Роршаха. Одна из самых популярных проективных методик связана с использованием чернильных пятен Роршаха (Aronow, & Reznikoff, 1983; Aronow, Reznikoff, & Moreland, 1994; Erdberg, & Exner, 1984; Exner, 1993). Разработанная швейцарским психиатром Германом Роршахом (Rorschach, 1921/1942), эта методика впервые была описана в 1921 г. Хотя стандартизованные серии чернильных пятен и раньше использовались психологами для изучения воображения и других функций, Роршах был первым, кто применил чернильные пятна для диагностического исследования лич-

[1] Класс теорий, развиваемых в рамках перцептуализма — учения, ярким представителем которого был американский философ Гамильтон. Центральная идея перцептуализма: поведение человека в каждый момент времени определяется текущим восприятием. — *Примеч. науч. ред.*

ности в целом. Развивая этот метод, Г. Роршах экспериментировал с большим количеством пятен, которые он предъявлял различным группам психически больных. В результате подобных клинических наблюдений те характеристики ответов, которые позволяли дифференцировать различные психиатрические синдромы, постепенно объединялись в систему показателей. Процедуры получения показателей затем оттачивались на основе дополнительного тестирования умственно отсталых и нормальных людей, а также художников, ученых и других четко различаемых профессиональных групп. Методология Роршаха, таким образом, представляла собой раннее, нестрогое и относительно субъективное применение метода привязки к внешнему критерию.

Вследствие безвременной смерти Роршаха, наступившей в 1922 г., разработку того, что было, в сущности, находящимся в работе тестом, продолжили его коллеги и ученики. В последующем десятилетии использование методики Роршаха существенно расширилось как в Европе, так и в Соединенных Штатах. Однако, из-за отсутствия единственного систематизатора, процедуры проведения, обработки и интерпретации результатов «подлинного Роршаха» стали быстро разрастаться и развиваться в отдельные методы или системы.[1] Начиная с 1960-х гг., говорить о тесте Роршаха как о едином, стандартизованном тесте было бы неверно. Различные системы и разных пользователей объединяют разве что 10 оригинальных стимульных таблиц (*cards*) и основные постулаты интерпретации, извлеченные из оригинальной работы Роршаха.

В тесте Роршаха используется 10 таблиц, на каждой из которых отпечатано двустороннее симметричное пятно, подобное изображенному на рис. 15–1. Пять пятен выполнены только в серо-черных тонах; два содержат дополнительные штрихи ярко-красного цвета, а остальные три сочетают цвета пастельных тонов. Типичная процедура проведения теста Роршаха сводится к следующему: респонденту показывают все таблицы, по одной за один раз, и просят рассказать о том, что могло бы изображать пятно на каждой из них. Помимо дословной записи ответов на каждую таблицу, проводящий обследование обычно отмечает время реакции и продолжительность ответа, положение или положения, в которых рассматриваются таблицы, непроизвольные реплики, эмоциональные проявления и другое побочное поведение респондента во время сеанса тестирования. После предъявления всех 10 карточек большинство пользующихся тестом Роршаха по определенной системе опрашивают обследуемого относительно частей и вида каждого из пятен, по которым возникали ассоциации. Во время этого опроса респонденты имеют к тому же возможность развить или пояснить прежние свои ответы.

Основные расхождения между разнообразными системами Роршаха, быстро разраставшимися в период с 1930-х по 1960-е гг., заключались в их методах обработки результатов и, следовательно, в вопросах интерпретации. По существу, предметом интереса при интерпретации результатов теста Роршаха можно сделать либо содержание ответов, либо формальные характеристики, наподобие локализации, детерминант, качества формы и разных количественных сводок, выводимые из ответов. Хотя системы Роршаха сильно различались деталями оценки и интерпретации ответов, на самом деле многие из них использовали общую базисную классификацию оценочных категорий. *Локализация (Location)* указывает на ту часть пятна, с которой респондент связывает свой ответ: используется ли при ответах все пятно, какая-то обычная деталь,

[1] Более подробное изложение истории создания теста Роршаха и его эволюции см. в Exner (1969, 1993).

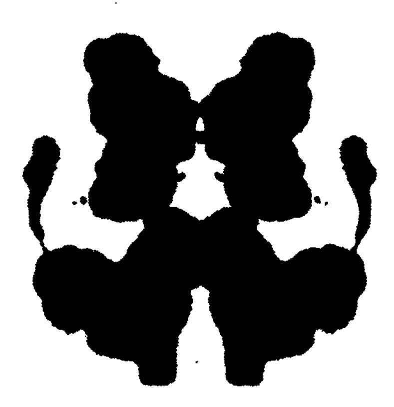


Рис. 15-1. Чернильное пятно типа используемых в методике Роршаха

необычная деталь, белая область или какая-то комбинация белого и темного участков. *Детерминанты (Determinants)* ответа включают форму, цвет, светотени (*shading*) и «движение». Хотя, конечно, в самом по себе чернильном пятне нет никакого движения, все же восприятие респондентом пятна как изображения движущегося объекта, относится к данной категории. Внутри этих категорий проводится более подробная дифференциация. Например, движение человека, движение животного, абстрактное движение или движение неодушевленного объекта подсчитываются отдельно. Аналогично, оттенки могут восприниматься как изображающие глубину, текстуру, неясные формы, такие как облака или как ахроматические репродукции цвета. *Качество формы (form quality)* или *уровень формы (form level)* ответов может относиться к точности, с какой они соответствуют использованной локализации, к их оригинальности или к тому и другому. В добавление к этому, в некоторых системах учитываются также когнитивная сложность ответов и другие качественные аспекты возникающих у респондента образов.

Трактовка *содержания (content)* тоже меняется в зависимости от системы оценки ответов, хотя некоторые основные категории используются постоянно. Главными среди них являются человеческие фигуры и их детали (или части человеческих фигур), фигуры животных и их части. К другим широко применяемым оценочным категориям можно отнести предметы, созданные людьми, растения, географические карты, облака, пятна крови, рентгеновские снимки, одежду, сексуальные объекты, ландшафты (пейзажи). На основе относительной частоты различных ответов, встречающихся у людей вообще,[1] часто определяется показатель популярности. По каждой из 10 таблиц определенные ответы относят к категории популярных ввиду их широкой встречаемости. Вдобавок ко всему, в большинстве систем Роршаха ведется учет (и подсчет) необычных или девиантных вербализаций со стороны обследуемого в процессе тестирования; такие вербализации особенно полезны при выявлении тяжелых форм психопатологии.

---

[1] «Люди вообще» — крайне расплывчатая характеристика. Роршах, например, под популярными ответами подразумевал толкования, которые даются каждым третьим респондентом, тогда как большинство авторов систем оценки ответов по тесту Роршаха относят к популярным ответы, встречающиеся у каждого шестого испытуемого (см., например, *Белый Б. И.* Тест Роршаха: Практика и теория / Под ред. Л. Н. Собчик. — СПб.: Дорваль, 1992. — С. 32–33). — *Примеч. науч. ред.*

Дальнейший анализ ответов в тесте Роршаха обычно основывался на относительном числе ответов, попадающих в различные категории, а также на определенных соотношениях и взаимосвязях между различными категориями. Типичные примеры качественных интерпретаций, применявшихся к ответам по тесту Роршаха, включает связывание «целостных» (*whole*) ответов с концептуальным мышлением, «цветовых» (*color*) ответов — с эмоциональностью, ответов типа «человеческое движение» (*human movement*) — с воображением и фантазией (миром иллюзий). При обычном применении теста Роршаха большое значение придается итоговой глобальной характеристике индивидуума, в которой клиницистом объединяются результаты различных частей протокола и учитываются взаимосвязи различных показателей и индексов. В реальной практике сведения, полученные из внешних источников, таких как другие тесты, беседы с пациентом и история болезни, также используются при подготовке этих характеристик.

Комплексная система Экснера. К 1960-м гг. тест Роршаха как психометрический инструмент получил дурную славу. Исследователи столкнулись с серьезными трудностями, коренящимися в самом методе Роршаха, а именно: с непостоянством суммарного количества ответов, влиянием эффектов тестирующего, взаимозависимостью показателей, а также с изобилием расходящихся систем оценки результатов. Эти обстоятельства превратили изучение надежности и валидности теста Роршаха в набор несогласованных, перегруженных методологическими просчетами и, в конечном итоге, неутешительных по своим результатам исследований. Многие клинические психологи продолжали систематически применять тест Роршаха, но большинство из них признавались, что не придерживаются строго какой-то одной системы оценки результатов. Вместо этого они предпочитали использовать данные теста Роршаха по своему усмотрению, начиная от полностью импрессионистской, качественной интерпретации и кончая более или менее строгим следованием одной или большему числу систем, которые они считали подходящими.

Многочисленные различия между пятью основными системами оценки ответов по тесту Роршаха, используемыми в США, были документально зафиксированы Джоном Экснером-младшим (Exner, 1969), работавшим вместе с Сэмюелем Беком (Samuel Beck) и Бруно Клопфером (Bruno Klopfer) — двумя из создателей наиболее расходящихся систем Роршаха.[1] В результате своих обширных исследований клинического применения теста Роршаха и анализа соответствующих литературных источников Экснер заинтересовался возможностью «дистилляции» в одну-единственную систему всех эмпирически обоснованных и полезных характеристик, которыми только мог обладать метод Роршаха. В течение последней четверти XX в. он предпринял самую дерзкую, но и оказавшуюся самой плодотворной из всех когда-либо предпринимаемых, попытку поставить тест Роршаха на психометрически прочную основу (Exner, 1974, 1991, 1993, 1995; Exner, & Weiner, 1995).

В общем и целом, Экснер разработал комплексную систему Роршаха, объединившую элементы, отобранные из пяти основных подходов к оценке результатов соответствующего теста. В этой комплексной системе (*Comprehensive System*) Экснер предлагает стандартизованные процедуры проведения, подсчета показателей и интерпрета-

[1] Создателями остальных трех систем были Маргарита Герц (Marguerite Hertz), Зигмунт Пиотровски (Zygmunt Piotrowski) и Дэвид Рапапорт (David Rapaport) с Роем Шейфером (Roy Schafer).

ции, отобранные на основе эмпирических сравнений разнообразных способов. Главное значение здесь придается не содержанию, а структурным переменным. Действительно, согласно Экснеру, цель обработки ответов — получение *структурной сводки (structural summary)*, которая образует ядро его системы и служит основой для большинства интерпретационных постулатов. Каждый ответ кодируется в соответствии с целым рядом различных оценочных категорий, включая, помимо прочих, локализацию, детерминанты, качество формы, содержание, организационную активность и популярность. Затем формируется список закодированных ответов и подсчитываются частоты кодов, после чего эти элементы используются при вычислении коэффициентов, процентных отношений и индексов, которые и составляют структурную сводку. Интерпретирующие утверждения могут выводиться, при наличии протокола полного теста Роршаха, из переменных на разных уровнях сложности. Одни гипотезы увязываются с простыми частотами, наподобие степени использования одной-единственной детерминанты (например, светотени); другие основываются на сочетанном проявлении двух или большего числа переменных, таких, например, как количество ответов с «человеческим» и «животным» содержанием. Самый сложный уровень анализа предполагает использование констелляций нескольких переменных и эмпирически выведенных критических показателей. Эти переменные группируются в индексы, — например, Индекс шизофрении (*Schizophrenia Index*), Индекс депрессии (*Depression Index*) и Индекс дефицита совладания (*Coping Deficit Index*), — которые, предположительно, отражают вероятность наличия определенных расстройств или состояний.[1]

Используя эту унифицированную систему, которая развивалась и совершенствовалась на протяжении двух десятков лет, Экснер и его единомышленники собрали внушительное количество психометрических данных, включая нормы для взрослых, детей и подростков, а также для различных психиатрических эталонных выборок. Исследования ретестовой надежности на различных временны́х интервалах, варьирующих от нескольких дней до трех лет, свидетельствуют о значительной временно́й устойчивости большинства оцениваемых переменных. Тщательно разработанные и не менее тщательно прописанные в комплексной системе Экснера руководящие принципы оценки ответов сделали возможным для подготовленных тестирующих добиваться довольно высоких коэффициентов согласия. Фактически, одним из самых важных вкладов работы Экснера является предоставление пользователям унифицированной системы Роршаха, допускающей сопоставимость данных разных специалистов, проводящих обследование по этой методике.[2] Поэтому вряд ли стоит удивляться, что комплексная система Экснера стала самым часто преподаваемым подходом к обработке и интерпретации теста Роршаха и оказалась полезной в том, что касается повышения статистической мощности исследований с применением методики Роршаха (см., например, Acklin, McDowell, & Orndoff, 1992; Ritzler, & Alter, 1986).

Несмотря на очевидные методологические усовершенствования, которые система Экснера внесла в методику Роршаха, несколько важных вопросов остаются нерешен-

---

[1] Имеются автоматизированные (компьютерные) программы-помощники для подсчета и интерпретации показателей в комплексной системе Экснера, которые требуют от пользователя только первоначального кодирования ответов.

[2] Однако вопрос о наилучшем способе измерения согласия оценщиков ответов по тесту Роршаха пока еще не решен. Обсуждение этой проблемы и сравнение трех различных методов оценивания надежности оценщика применительно к данным теста Роршаха см. в работе McDowell, & Acklin (1996).

ными. Главный и наиболее сложный из них — вопрос валидности этого инструмента. Литература по данной теме поистине огромна и до сих пор полна взаимоисключающих выводов. Тест Роршаха, как и *MMPI*, использовался в самых широких целях, многие из которых явно выходили за пределы намерений его автора, и эта множественность областей применения еще больше осложнила исследование валидности данного теста. В общем, метааналитическое исследование показало, что индексы конвергентной валидности для теста Роршаха сопоставимы с таковыми для *MMPI* (Atkinson, Quarrington, Alp, & Cyr, 1986; K. C. H. Parker, Hanson, & Hunsley, 1988). В добавление к этому, результаты большой работы с тестом Роршаха, проделанной самим Экснером, обеспечивают существенную поддержку валидности многих конструктов, оцениваемых посредством его комплексной системы, и способствуют сохранению полезности этой системы для описания определенных аспектов функционирования личности. Тем не менее когда разработанная Экснером комплексная система Роршаха начинает применяться для диагностики сложных текущих состояний или предсказания будущего поведения, результаты исследований оказываются неоднородными (Exner, 1996; Weiner, 1994a; Wood, Nezworski, & Stejskal, 1996a, 1996b).

Основным затрудняющим интерпретацию показателей теста Роршаха фактором является суммарное число ответов, известное как продуктивность ответов, или *R*. Когда отдельные люди или группы существенно различаются между собой по величине *R*, это может привести к появлению различий по другим оценочным категориям. Таким образом, установленные в определенных категориях различия могут оказаться всего лишь артефактом вариаций суммарного числа ответов. Эту характерную особенность показателей теста Роршаха может усугублять еще и то, что продуктивность ответов, по-видимому, связана с другими переменными, такими как интеллектуальный уровень и объем образования.[1] Характерно, что специалисты, работающие с тестом Роршаха, различаются своим отношением к важности и сложности проблем, создаваемых вариациями в продуктивности ответов. Мейер (G. J. Meyer, 1992, 1993), например, призывал к более глубокому изучению психологического смысла *R* и сравнительному исследованию приобретений и потерь в случае контролирования или регулирования этой переменной. Другие считают, что уже достаточно доказательств незначимого воздействия данной переменной в большинстве ситуаций (Exner, 1992; Weiner, 1995b). Есть и такие, кто утверждает, что к проблемам, создаваемым переменной *R*, нужно подходить по-разному, в зависимости от области применения и назначения полученных данных, а в случае научных исследований — в зависимости от степени отклонения распределений от нормального (Kinder, 1992; Lipgar, 1992).

Разработанная Экснером комплексная система не осталась без внимания критики. Многие пользователи теста Роршаха выступили против чисто эмпирического подхода Экснера и низкого уровня использования содержательных данных, что, по их мнению, существенно снижает клиническую полезность этого теста. Критики также отмечали чрезмерную сложность системы Экснера наряду с ее эпизодически проявляющейся расплывчатостью и противоречивостью. Кроме того, в собственных исследованиях Экснера было найдено много недостатков, включая малый объем выборок, большое

[1] Что касается показателя *R*, в исследовательской литературе, посвященной выяснению возможностей симуляции результатов теста Роршаха, приводится интересный результат: когда испытуемых просят симулировать психоз по показателям Роршаха, происходит снижение продуктивности ответов (G. G. Perry, & Kinder, 1990).

число переменных, нехватку исследований по кросс-валидизации и недоступность широкой проверки правильности полученных результатов.[1] Как бы то ни было, остается фактом, что наличие системы Экснера, вместе с собранными им и его коллегами исследовательскими данными, вдохнуло новую жизнь в методику Роршаха как психометрический инструмент.

Альтернативные подходы. Несмотря на широкое признание комплексной системы Экснера, наряду с ней существует целый ряд совершенно иных подходов к тесту Роршаха. Фактически, многие из тех, кто поддержал усилия Экснера по возрождению интереса к методике Роршаха, стояли именно на этих иных позициях. Один из таких альтернативных подходов, с гораздо более выраженной клинической ориентацией, описан Ароновым и его коллегами (Aronow, & Reznikoff, 1976, 1983; Aronow et al., 1994, 1995). Этот подход трактует тест Роршаха по существу как стандартизованное клиническое интервью, которое нацелено на выборочное обследование перцептивных операций конкретного человека. Таким образом, в рамках этого подхода главное внимание уделяется интерпретации содержания, а не структурным переменным или перцептивным детерминантам ответов респондента. Тем не менее имеющиеся шкалы содержания и системы подсчета показателей в плане психометрических требований не считаются достаточно надежными для того, чтобы использоваться в индивидуальной диагностике. Скорее, авторы данного подхода рекомендуют строго клиническое применение методики Роршаха как средства улучшения идиографического понимания каждого конкретного случая, и они отмечают, что большинство опытных клиницистов тяготеют к этому подходу вследствие его полезности в процессе психотерапии. Их интерпретации опираются на содержание ответов, дополненное вербальным и невербальным поведением респондента в ходе тестирования. На основе опубликованных исследований и клинического опыта Аронов и его сотрудники подготовили комплект методических рекомендаций для получения более эффективных и надежных идиографических интерпретаций. Например, они указывают на то, что ответы, которые выходят за границы банальностей и менее ограничены стимульными свойствами конкретных пятен, с большей вероятностью будут нести в себе значимую для понимания индивидуального случая информацию. Точно так же эти авторы предостерегают от применения жестких систем символической интерпретации, которые придают категориям содержания фиксированные значения или приписывают пятнам Роршаха способность вызывать в памяти всегда одни и те же образы. Вместо этого они предлагают процедуры для выявления смыслового значения ответов, отличающиеся довольно консервативной манерой, совместимой с общими психодинамическими принципами, и использующие сведения об истории жизни конкретного человека.

Другой современной альтернативой системе Экснера является подход, поддерживаемый Полом Лернером (P. M. Lerner, 1991). Работа Лернера представляет собой крайнюю противоположность эмпирическому (*atheoretical*) образу мышления Экснера в том смысле, что она насквозь пропитана духом современной психоаналитической теории, в рамках которой Лернер начал в 1970-х гг. развивать свою систему. Тогда как

---

[1] Выборочно познакомиться с критическими замечаниями в адрес работы Экснера, включая вышеупомянутые, можно по следующим публикациям: Aronow, Reznikoff, & Moreland (1995), Kleiger (1992), P. M. Lerner (1994), W. Perry (1993), Viglione (1989), Vincent, & Harman (1991), Wood, Nezworski, & Stejskal (1996a, 1996b).

Экснер (Exner, 1989) утверждает, что методика Роршаха — это тест, в котором проекция редко вступает в действие, Лернер считает тест Роршаха по сути своей проективным методом для оценки внутреннего мира индивидуума.[1] Руководство Лернера снабжает пользователей теста Роршаха ориентирами для его применения в клинических и исследовательских целях при решении задач оценивания репрезентаций объекта (*object representations*), защитных маневров (*defensive maneuvers*) и других основных понятий современных психодинамических теорий.

Особый случай клинического применения методики Роршаха иллюстрируется так называемым «консенсусом по Роршаху» (*the consensus Rorschach* — Aronow et al., 1994; chap. 13; Blanchard, 1968; Cutter, & Farberow, 1970). В этой адаптации теста Роршаха чернильные пятна предъявляются для совместной интерпретации супружеским парам или другим членам семьи, сослуживцам, членам молодежных группировок или другим естественным группам. Посредством обсуждения и согласования позиций участники должны прийти к единой, общей совокупности ответов. Методика использовалась, с несомненным успехом, в качестве опорного элемента при изучении межличностных отношений и других разновидностей социального поведения.

Методика Роршаха была исключительно точно охарактеризована как «тест, который неоднократно пережил свои похороны» (Peterson, 1994, p. 396). Его кончину пророчили много раз, потому что, подобно всем наиболее часто используемым психологическим тестам, тест Роршаха оказывался также в числе тестов, которые чаще всех других применялись не по назначению. В настоящее время эта методика переживает период чудесного воскрешения в плане исследовательской активности и теоретического обоснования. И хотя продолжают существовать различные подходы в работе с тестом Роршаха, пользователи всех рангов, кажется, осознали, наконец, что он представляет особую ценность при изучении перцептивных, когнитивных и аффективных аспектов функционирования личности. Одних удовлетворяет отношение к тесту Роршаха просто как к методу сбора данных, который можно использовать занимая разные теоретические позиции, тогда как другие тратят много сил на то, чтобы связать эмпирическую и теоретическую традиции, а заодно и различные теоретические подходы к методике Роршаха, в полностью интегрированную, комплексную систему (см., например, Acklin, 1995; Blatt, 1990; P. M. Lerner, 1994; Meloy, & Singer, 1991; Weiner, 1994b; Willock, 1992).

Методика чернильных пятен Хольцмана. Серьезная попытка придать методике чернильных пятен психометрическую направленность была предпринята Уэйном Хольцманом (Wayne H. Holtzman) еще до того как Экснер начал разрабатывать свою комплексную систему. Сделанная по образцу теста Роршаха, методика чернильных пятен Хольцмана (*Holtzman Inkblot Technique* [*HIT*]) конструировалась таким образом, чтобы исключить главные технические недостатки своего прототипа (Holtzman, 1961, 1986; Holtzman, Thorpe, Swartz, & Herron, 1961). Впрочем, изменения в стимульных материалах и процедуре достаточно обширны, чтобы считать тест Хольцмана новым тестом и оценивать его безотносительно к тесту Роршаха. Методика Хольцмана предусматривает две параллельные серии таблиц (*cards*), по 45 таблиц каждая; чернильные пятна отбирались из большой черновой совокупности на основе эмпирических критериев, нацеленных на максимизацию их эффективности. По каждой карточке

---

[1] Причины такого явного расхождения позиций кроются в различии взглядов авторов на природу проекции и задачу респондента в тесте Роршаха.

получают только один ответ. Имеются как ахроматические, так и цветные таблицы; несколько чернильных пятен имеют выраженную асимметрию.

Проведение и подсчет показателей HIT хорошо стандартизованы и ясно описаны с самого начала. Показатели получают по 22 параметрам ответов, среди которых многие аналогичны параметрам или, иначе говоря, переменным теста Роршаха, и есть ряд дополнительных переменных, таких как тревожность и враждебность. По каждой переменной имеются процентильные показатели для нормальных выборок детей и взрослых, а также для ряда девиантных групп (E. F. Hill, 1972; Holtzman, 1975). Надежность оценщика выглядит весьма удовлетворительно. Исследования надежности эквивалентных половин и взаимозаменяемых форм теста, а также ретестовой надежности обнаружили различия в отношении разных параметров ответов, хотя большинство результатов оказались обнадеживающими. Групповая форма этого теста, в которой используются слайды, дает показатели по большинству переменных, сопоставимые с показателями, получаемыми при индивидуальном проведении теста (Holtzman, Moseley, Reinehr, & Abbott, 1963; Swartz, & Holtzman, 1963). *HIT 25* — краткая версия, состоящая из первых 25 таблиц *HIT* (*Form A*) и предполагающая получение двух ответов по каждой таблице, была недавно предложена Хольцманом (Holtzman, 1988) и находится в процессе нормирования (Swartz, 1992).

Накоплен значительный объем данных по валидности HIT, большая часть которых выглядит многообещающе (Gamble, 1972; Holtzman, 1975, 1986, 1988; Leichsenring, 1991; Sacchi, & Richaud de Minzi, 1989; Swartz, 1973). В исследованиях валидности применялись разнообразные подходы, включая изучение тенденций возрастного развития, кросс-культурные сравнения, корреляции с другими тестами и с поведенческими признаками особенностей личности, а также сравнение контрастных групп, образованных как из нормальных респондентов, так и из психически больных лиц. Подготовленный Хиллом (E. F. Hill, 1972) справочник по применению *HIT* имеет сугубо клиническую ориентацию.

Очевидно, что с психометрической точки зрения *HIT* имеет ряд преимуществ перед тестом Роршаха. Наличие параллельных форм позволяет не только определить ретестовую надежность, но и проводить контрольные исследования (*follow-up studies*). Ограничение ответов до одного на каждую таблицу делает продуктивность ответов (R) постоянной для каждого респондента, что позволяет избежать многих недостатков подсчета показателей в тесте Роршаха. Нужно отметить, однако, что длина ответа (количество слов) в этом тесте пока еще не контролируется и, как в тесте Роршаха, оказалась значимо связанной с некоторыми показателями *HIT* (Megargee, 1966). И все же, несмотря на некоторые преимущества *HIT*, о нем накоплено значительно меньше информации, чем по тесту Роршаха, и нужны дополнительные данные для установления диагностической значимости многих его показателей и конструктной валидности оцениваемых им параметров личности (что касается рецензий, см. Cundick, 1985; Dush, 1985).

Рисуночные методики

Тест тематической апперцепции. В противоположность методике чернильных пятен Тест тематической апперцепции (*Thematic Apperception Test* [*TAT*]) предлагает обследуемому гораздо более структурированные стимулы и требует от него более сложных и более организованных в смысловом плане словесных ответов. Интерпрета-

ция ответов проводящим обследование обычно основывается на преимущественно качественном контент-анализе. Разработанный Генри Мюрреем и его сотрудниками (Murray et al., 1938) в Гарвардской психологической клинике (*Harvard Psychological Clinic*), Тест тематической апперцепции[1] не только широко использовался в клинической практике и исследованиях, но и послужил образцом для разработки многих других инструментов (J. W. Atkinson, 1958; Bellak, 1993; Dana, 1996b; R. Harrison, 1965; Holmstrom, Silber, & Karp, 1990; Klopfer & Taulbee, 1976, p. 554–558; Obrzut, & Boliek, 1986).

Материалы *TAT* состоят из 19 таблиц (*cards*) с расплывчатыми черно-белыми картинками и одной пустой таблицы.[2] Респондента просят сочинить по каждой картинке историю о том, что привело к изображенному на ней событию, что происходит в данный момент, что чувствуют и думают действующие лица и чем все это закончится. В случае предъявления пустой таблицы респонденту дается инструкция вообразить какую-либо картину на этой таблице, описать ее, а затем рассказать о ней историю. Оригинальная процедура, описанная Г. Мюрреем в руководстве к *TAT*, требует двух часовых сеансов, в каждом из которых используется по 10 таблиц. Предназначенные для второго сеанса таблицы сознательно подбирались так, чтобы изображенные на них картины выглядели более необычными, драматическими и странными, чем картинки на таблицах для первого сеанса, а сопровождающие их инструкции побуждали респондентов дать волю своему воображению. Существуют четыре частично совпадающих набора по 20 таблиц каждый: для мальчиков, девочек, лиц мужского пола старше 14 лет и лиц женского пола старше 14 лет. Большинство клиницистов пользуются сокращенными наборами специально отбираемых таблиц, редко предъявляя более 10 таблиц одному респонденту.

Следуя оригинальному методу интерпретации историй *TAT* (Murray et al., 1943), проводящий обследование сначала определяет, кто является «героем» — персонажем того или иного пола, с которым респондент предположительно идентифицирует себя. Затем анализируется содержание историй исходя, главным образом, из перечня «потребностей» (*needs*) и видов «давления» (*press*), составленного Мюрреем. Некоторые из приводимых в нем потребностей были описаны в главе 13, в связи с рассмотрением Списка личных предпочтений Эдвардса (*EPPS*). Примеры включают потребности достижения, аффилиации и агрессии. *Давление* относится к воздействиям среды, которые могут облегчать или затруднять удовлетворение потребностей. Когда человек подвергается нападкам или критике, испытывает проявления чьей-то любви, ощущает чью-то поддержку или чувствует незащищенность перед физической опасностью, — все это иллюстрации давления. При оценке важности или силы конкретной потребности или давления для индивидуума особое внимание уделяется интенсивности, длительности и частоте их появления в различных историях, а также своеобразию их связи с данной картинкой. Предполагается, что необычный материал, отличающийся от стереотипных ответов на картинку, по всей вероятности, более значим для индивидуума.

---

[1] Хотя такой перевод названия этого теста прижился в отечественной психологии, вполне допустимым (а с моей точки зрения, и более адекватным) является перевод «тематический тест апперцепции». — *Примеч. науч. ред.*

[2] Изложение увлекательной истории происхождения используемых в *TAT* рисунков можно прочитать в статье W. G. Morgan (1995).

Опубликовано довольно много нормативной информации относительно характеристик наиболее частых ответов по каждой таблице, включая манеру их восприятия, развиваемые темы, приписываемые персонажам роли, эмоциональный тон и скорость ответов, длину историй и т. д. (J. W. Atkinson, 1958; W. E. Henry, 1956; Murstein, 1972). Хотя эти нормативные данные обеспечивают общую основу для интерпретации индивидуальных ответов, большинство клиницистов больше полагаются на «субъективные нормы», построенные на собственном опыте работы с тестом и на информации, которую они собрали об обследуемом человеке с помощью других средств. Был также разработан ряд рейтинговых шкал и схем количественной оценки ответов по *TAT*, которые дают хорошие коэффициенты надежности оценщика. Однако вследствие того, что их применение связано с большими затратами времени, такие оценочные процедуры редко используются в клинической практике. Типично предъявляемый в клинической ситуации как индивидуальный устный тест, *TAT* может также проводиться в письменной форме и как групповой тест.

TAT широко использовался в исследованиях личности. К сожалению, большое разнообразие процедур проведения и обработки результатов, да и стимульного материала тоже, скрывающееся под заголовком *TAT*, обнаружило себя не только в клинической практике, но и в научных исследованиях (Keiser, & Prather, 1990). Это разнообразие сильно затруднило изучение психометрических свойств «*TAT* как отличного от других психологического теста». В добавление к этому получено достаточно много экспериментальных данных, которые показывают, что такие состояния, как голод, недосыпание и социальная фрустрация, существенно влияют на характер ответов по *TAT* (J. W. Atkinson, 1958). Хотя эти данные поддерживают «проективную гипотезу», чувствительность *TAT* к таким временным состояниям может вносить путаницу в смысл ответов. Вопрос внутренней согласованности ответов по *TAT* также не остался без внимания исследователей (J. W. Atkinson, 1981; Entwisle, 1972). Кроме того, следует в известной степени контролировать длину историй, или продуктивность, — проблема, общая у *TAT* с тестом Роршаха (J. W. Atkinson & Raynor, 1974, chap. 3).

Тем не менее ценность методик тематической апперцепции вообще, и *TAT* в частности, не подвергается сомнению. Недавние исследования вновь подтвердили клиническую полезность различных версий *TAT* как для традиционных приложений, таких как оценивание степени психопатологии и использования защитных механизмов, так и для решения новых задач, таких как оценка навыков решения проблем (Cramer, & Blatt, 1990; Hibbard et al., 1994; Ronan, Colavito, & Hammontree, 1993; Ronan, Date, & Weisbrod, 1995). Одно из самых многообещающих приложений *TAT* связывают с недавно разработанными шкалами для клинической оценки объектных отношений (Alvarado, 1994; Barends, Westen, Leigh, Silbert, & Byers, 1990; Freedenfeld, Ornduff, & Kelsey, 1995; Westen, 1991; Westen, Lohr, Silk, Gold, & Kerber, 1990). Полезность *TAT* ни коим образом не ограничивается тематическим анализом ответов; не менее плодотворным при изучении отдельных людей и групп может быть использование формальных характеристик структуры и содержания придуманных респондентами историй (см., например, Cramer, 1996; McGrew, & Teglasi, 1990; Teglasi, 1993).

Адаптации *TAT* и родственные тесты. Многие адаптации *TAT* были разработаны для особых целей и потому обнаруживают разную степень сходства с оригиналом. Вопрос о том, где провести границу между модифицированными версиями *TAT* и новыми тестами, основанными на том же общем подходе, что и *TAT*, решается произвольно. Несколько версий *TAT* было подготовлено для использования в исследованиях

аттитюдов в отношении проблем рабочих, национальных меньшинств, власти и т. д. (D. T. Campbell, 1950; R. Harrison, 1965). Другие адаптации разрабатывались для использования в профконсультировании, аттестации управленческих кадров и в разнообразных исследовательских проектах. Кроме того, были сконструированы формы ТАТ для специфических популяций, в том числе дошкольников, младших школьников, детей-инвалидов, подростков, а также различных национальных и этнических групп (R. Harrison, 1965).

Некоторые из адаптаций *ТАТ* предназначены для тщательного измерения одной-единственной потребности или побуждения, например сексуальных отношений (*n-Sex*) или агрессии (*n-Agg*). Особый интерес представляет широкое изучение потребности достижения (*n-Ach*), проводившееся на протяжении 30 лет МакКлелландом, Аткинсоном и их сотрудниками (J. W. Atkinson, 1958; J. W. Atkinson, & Feather, 1966; J. W. Atkinson, & Raynor, 1974; McClelland, 1985; McClelland, Atkinson, Clark, & Lowell, 1953/1976). Для измерения *n-Ach* применяются четыре картинки, две из которых взяты из *ТАТ.* Были разработаны подробные схемы для обработки придуманных респондентами историй с целью определения выраженности *n-Ach.* Эта методика использовалась в обширной программе исследований мотивации достижения. Диапазон изучавшихся проблем был весьма широк: от базисной теории мотивации (J. W. Atkinson, & Feather, 1966) до социальных источников и последствий n-Ach и ее роли в расцвете и упадке обществ (McClelland, 1961/1976). Метаанализ исследований, сравнивающих *ТАТ* и опросники для измерения *n-Ach*, свидетельствует о том, что оба этих метода являются валидными, хотя и применяются с различными целями и для оценки разных аспектов стремления к достижениям (Spangler, 1992).

Краткий справочник по системам количественных показателей, используемым в контент-анализе вербального материала, был подготовлен в 1992 г. Чарльзом Смитом (Charles Smith) при участии Джона Аткинсона (John W. Atkinson), Дэвида МакКлелланда (David C. McClelland) и Джозефа Вероффа (Joseph Veroff). Наряду с системами, завоевавшими прочную репутацию в традиционных исследованиях, посвященных, скажем, измерению мотивов достижения, аффилиации и власти, в этот справочник включено множество других систем количественных показателей, имеющих отношение к столь разным темам, как политическая идеология и пределы самообладания. Рассматриваются также некоторые концептуальные вопросы и методологические соображения по поводу взятия выборок, количественной оценки и анализа вербального материала. Хотя многие из описанных в этом справочнике систем основаны на модификациях *ТАТ* и отражают взгляды Г. Мюррея, в нем представлен и ряд других теоретических перспектив. Более того, включенные в него системы предназначены для анализа явного, а не символического содержания выборочных образцов мышления, — в отличие от «проекций», — а также для исследовательских целей, а не для клинического использования.

Хотя утверждается, что оригинальный *ТАТ* применим к детям, достигшим 4-летнего возраста, для детей от 3 до 10 лет специально разработан Тест детской апперцепции (*Children's Apperception Test* [*CAT*] — Bellak, 1993). На таблицах *CAT* изображения людей заменены изображениями животных исходя из предположения, что маленьким детям легче осуществлять проекцию на изображения животных, чем людей[1]. Раз-

---

[1] Что касается конкретной информации о клиническом применении *ТАТ* и других методик «придумывания историй» (*storytelling techniques*) в работе с детьми, см. Teglasi (1993) и Worchel, & Dupree (1990).

нообразные животные в рисунках *CAT* изображены в типично человеческих ситуациях, в антропоморфной манере, характерной для комиксов и детских книг. Картинки предназначены пробудить фантазии, связанные с проблемами кормления и другой оральной активностью, соперничеством сиблингов, взаимоотношениями родителей и детей, агрессией, приучением к туалету и с другими видами детского опыта. Авторы CAT подготовили модификацию своего теста (*CAT-H*)[1] с изображениями людей для использования с более взрослыми детьми, особенно с теми, чей умственный возраст превышает 10 лет (Bellak, & Hurvich, 1966). Они утверждают, что в зависимости от возраста и особенностей личности ребенка более эффективной может оказаться либо одна, либо другая форма *CAT* (что касается критических разборов CAT, см. Hatt, 1985; Shaffer, 1985).

Разработанный позднее Тест апперцепции для детей Робертса (*Roberts Apperception Test for Children* [*RATC*]) в большей степени отвечает психометрическим стандартам конструирования и оценки тестов, чем другие методики этого типа (McArthur, & Roberts, 1982; см. также Sines, 1985). *RATC* предусматривает два частично перекрывающихся набора по 16 стимульных таблиц (*cards*) в каждом: один для мальчиков, другой для девочек. Кроме того, имеется дополнительный набор с изображениями чернокожих детей, но по нему не были установлены нормы. Для этого теста отбирались картинки, изображающие семейные межличностные ситуации, в которых дети вступают в отношения со взрослыми или другими детьми (см. рис. 15–2). Придуманные детьми истории оцениваются по ряду шкал, охватывающих те типы проблем, из-за которых детей обычно приводят к клиническим психологам. Ясные и подробные методические указания к этому тесту позволяют проводить достаточно объективную оценку детских ответов. Нормы основаны на ответах 200 детей, названных учителями хорошо приспособленными к действительности (*well-adjusted*). Сравнение этих ответов с ответами 200 детей, наблюдаемых в медико-психологическом детском центре (*child guidance clinic*), позволило получить некоторые данные о валидности, приведенные в руководстве к *RATC*. Бесспорно, этот инструмент представляет серьезную попытку объединить гибкость проективных методик с процедурами проведения, обработки и оценки результатов стандартизованного теста. Исследования валидности *RATC* для разных областей использования продолжают давать благоприятные результаты (см., например, Palomares, Crowley, Worchel, Olson, & Rae, 1991). Вдобавок ко всему, Глен Робертс (G. E. Roberts, 1994) подготовил справочник, содержащий подробные методические рекомендации по обработке и интерпретации результатов *RATC* при его клиническом применении.

TEMAS — испанское слово, означающее «темы», и остроумный акроним для *Tell-Me-A-Story* («Расскажи мне историю») — инструмент, специально сконструированный для оценки когнитивных, аффективных и личностных характеристик детей в возрасте от 5 до 18 лет (Costantino, Malgady, & Rogler, 1988). В *TEMAS* используются два параллельных набора многоцветных стимульных таблиц (*cards*): один для детей этнических меньшинств, другой для белых детей. Стимульные материалы тщательно разрабатывались с тем, чтобы облегчить вербальную продукцию и стимулировать придумывание историй, связанных с выбором конфликтующих целей типа немедленное/отсроченное удовольствие. На картинках из набора для детей этнических меньшинств изображены персонажи, черты и цвет лица которых намекают на африканское или

[1] Прибавка *H* — первая буква слова *Human*. — *Примеч. науч. ред.*



Рис. 15–2. Одна из картинок, используемых в Тесте апперцепции для детей Робертса *(Copyright © 1982 by Western Psychological Services. Воспроизводится с разрешения)*

испанское происхождение. Несмотря на восхваление *TEMAS* как явно более совершенного инструмента по сравнению с таблицами оригинального *TAT* в смысле его пригодности для обследования афроамериканских и испаноамериканских детей, психометрические свойства этого инструмента, особенно его ретестовая надежность и внутренняя согласованность неоднократно подвергались сомнению (что касается критических разборов *TEMAS*, см. Dana, 1993, chap. 8; Lang, 1992; Ritzler, 1993a).

Сходные тесты тематической апперцепции были разработаны для пожилых людей, в том числе Геронтологический тест апперцепции (*Gerontological Apperception Test* — Wolk & Wolk, 1971) и Тест апперцепции пожилых людей (*Senior Apperception Test* — Bellak, 1993; Bellak & Bellak, 1973). В том и другом тесте используются наборы таблиц (cards), изображающих одного или более пожилых людей и иллюстрирующих проблемы, которые могут беспокоить стариков, например одиночество, семейные трудности, зависимость и немощность. Оба инструмента были раскритикованы за преждевременную публикацию и использование картинок, которые поддерживали «вредящие стереотипы старения» (J. P. Schaie, 1978; K. W. Schaie, 1978). К тому же ни один из них не обнаружил преимуществ перед *TAT* в тестировании пожилых людей (Fitzgerald, Pasewark, & Fleisher, 1974; Foote, & Kahn, 1979), и Геронтологический тест апперцепции больше не издается.

Анализ рисуночной фрустрации Розенцвейга. В Тесте тематической апперцепции и родственных ему методиках, которые мы только что рассмотрели, картинки используются для того, чтобы стимулировать свободную игру воображения и вызывать сложные вербальные ответы. В противоположность этому, Анализ рисуночной фрустрации Розенцвейга (*Rosenzweig Picture-Frustration Study* [*P-F study*]), описываемый



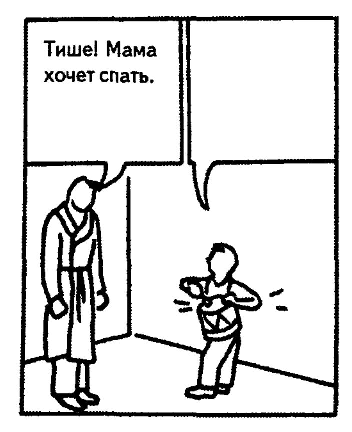


Рис. 15—3. Типичные задания из Анализа рисуночной фрустрации Розенцвейга (форма P-F Study для детей)
(Copyright © 1976 by Saul Rosenzweig. Воспроизводится с разрешения)

в этом разделе, дает меньше простора для фантазии и требует более простых ответов. Имеются в наличии три отдельных формы этого инструмента: для взрослых от 14 лет и старше (Rosenzweig, 1950, 1978a, 1978d), для подростков от 12 до 18 лет (Rosenzweig, 1970, 1976b, 1981a) и для детей от 4 до 13 лет (Rosenzweig, 1960, 1977, 1981b, 1988). Созданный на основе принадлежащей автору теории фрустрации и агрессии, *P-F Study* представляет собой серию условных рисунков, на которых один персонаж каким-то образом срывает («фрустрирует») намерения и действия другого или привлекает внимание к фрустрирующей ситуации. Два образца картинок из формы *P-F Study* для детей приведены на рис. 15–3. На специально отведенном пустом месте тестовой карточки респондент пишет, что ответил бы, по его мнению, фрустрированный персонаж.

Ответы классифицируются по типу и направлению агрессии. По типу агрессии различают препятственно-доминантные (*obstacle-dominance*) реакции, придающие особое значение фрустрирующему объекту, эго-защитные (*ego-defense*) реакции, фокусирующие внимание на защите фрустрированного субъекта, и потребностно-персистентные (*need-persistence*) реакции, концентрирующиеся на конструктивном решении фрустрирующей проблемы. По направлению агрессии реакции оценивают как: экстраагрессивные (*extraggressive*), или направленные на внешнее окружение; интраагрессивные (*intraggressive*), или обращенные на себя, и имагрессивные (*imaggressive*), в которых «агрессивная энергия» переводится в попытку замять проблему или выкрутиться из фрустрирующей ситуации.[1] При определении показателей теста процент ответов, попадающих в каждую из этих категорий, сравнивается с соответствующими нормативными процентами ответов. Можно также получить оценку уровня групповой конформности (*group conformity rating* [*GCR*]), отражающую тенденцию индивидуума давать ответы, согласующиеся с модальными ответами выборки стандартизации.

[1] Более привычные для отечественного читателя названия этих реакций: экстрапунитивные (*extrapunitive*), интропунитивные (*intropunitive*) и импунитивные (*impunitive*) соответственно. — *Примеч. науч. ред.*

Являясь более ограниченным по диапазону изучаемых реакций, более структурированным и относительно объективным в методиках подсчета показателей, *P-F Study* лучше поддается статистическому анализу, чем большинство других проективных методик. С самого начала прилагались систематические усилия по сбору норм и проверке надежности и валидности этого инструмента. За 50 лет с *P-F Study* его автором и другими специалистами было проведено большое количество исследований, результаты которых отражены в публикациях, посвященных, в основном, оценке психометрических свойств этого инструмента и таким темам, как клиническая диагностика, связанные с возрастным развитием изменения, половые различия, культурные различия и отношения между юмором и агрессией (Graybill, 1990, 1993; Nevo, & Nevo, 1983; Rosenzweig, 1976a, 1978b, 1978c; Rosenzweig, & Adelman, 1977; что касается рецензий *P-F Study*, см. Viglione, 1985; Wagner, 1985).

Вербальные методики

Хотя все обсуждавшиеся до сих пор проективные методики требуют вербальных ответов, среди методик этой группы есть и полностью вербальные, т. е. использующие слова и в качестве стимульного материала, и в ответах. Часть таких вербальных методик может применяться как в устной, так и в письменной форме, но все они пригодны для письменного предъявления при групповом тестировании. Конечно, в последнем случае предполагается минимальное умение читать и достаточно хорошее владение языком, на котором разрабатывался тест, а это значит, что подобные методики неприменимы к маленьким детям, неграмотным или говорящим на другом языке.

Методикой, опередившей волну проективных тестов более чем на полвека, является *тест словесных ассоциаций (word association test)*. Названная вначале «тестом свободных ассоциаций» (*free association test*), эта методика была впервые систематически описана Гальтоном (Galton, 1879). В. Вундт и Дж. М. Кэттелл впоследствии включили ее в инструментарий психологической лаборатории, приспособив для решения многих задач. Процедура проведения теста состоит в предъявлении испытуемому серии не связанных между собой слов, на каждое из которых его просят отвечать первым пришедшим на ум словом. Первые психологи-экспериментаторы, так же как и первые тестологи, работавшие с умственными тестами, видели в таких ассоциативных тестах средство для изучения процессов мышления.

Клиническое применение методов словесных ассоциаций было стимулировано, главным образом, психоаналитическим движением, хотя другие психиатры, такие как Э. Крепелин, и раньше проводили исследования с помощью подобных методик. Среди психоаналитиков наибольший вклад в систематическую разработку теста словесных ассоциаций принадлежит К.-Г. Юнгу. Юнг (Jung, 1910) отбирал слова-стимулы, символизирующие распространенные «эмоциональные комплексы», и анализировал ответы относительно времени реакции, содержания и телесных выражений эмоционального напряжения. Тридцать лет спустя аналогичная методика словесных ассоциаций была разработана в клинике Меннингера Дэвидом Рапапортом (D. Rapaport, 1946/1968). Согласно его авторам, тест имел двоякую цель: помочь в обнаружении нарушений процессов мышления и подсказать значимые области конфликта. Можно также упомянуть об использовании методики словесных ассоциаций как «детектора

лжи».[1] Инициатива применить ее для такой цели также принадлежит Юнгу, и впоследствии подобное применение методики словесных ассоциаций широко изучалась как в лабораторных условиях, так и в практических ситуациях (Burt, 1931; Lindsley, 1955). Обоснование, предложенное для оправдания применения словесных ассоциаций в целях выявления лжи или виновности, было похоже на обоснование использования этого метода для обнаружения областей эмоционального конфликта.

Иллюстрацией иного подхода к тесту словесных ассоциаций служит ранняя работа Кента и Розанова (Kent & Rosanoff, 1910). Разрабатывавшийся в основном как инструмент психиатрического скрининга, в Тесте свободных ассоциаций Кента—Розанова (*Kent-Rosanoff Free Association Test*) использована совершенно объективная система количественных показателей. В качестве слов-стимулов были взяты 100 общеупотребительных, нейтральных слов, отобранных в силу того, что они обычно вызывали у людей одни и те же ассоциации. Например, на слово *стол* большинство людей отвечают *стул*, на слово *темно* — *светло*. Был подготовлен комплект частотных таблиц — одна на каждое слово-стимул, — показывающих, сколько раз каждый ответ давался в выборке, составленной из 1000 «нормальных» взрослых людей. При оценке результатов теста Кента—Розанова на основе частот ответов каждого обследуемого выводился «индекс общности» (*index of commonality*). Сравнения психотиков с нормальными людьми говорили о том, что психотики получают более низкий индекс общности, чем нормальные.

Однако диагностическое использование теста снижалось по мере того, как становилось понятным, что частота ответов существенно варьирует в зависимости от возраста, социоэкономического и образовательного уровня, регионального и культурного происхождения, креативности и других факторов. Следовательно, правильная интерпретация показателей требует накопления норм по многим подгруппам и их периодического обновления, по мере введения в употребление новых и изменения частотности старых слов. Вдобавок, популярность традиционных психоаналитических понятий, которые стимулировали разработку этих методик, пошла на убыль (Rabin & Zlotogorski, 1981). Тест Кента—Розанова тем не менее сохранил свое положение стандартного лабораторного инструмента. В некоторых странах были собраны дополнительные нормы, а сама методика широко применялась в исследованиях вербального поведения и личности (Goldfarb, & Halpern, 1984; Isaacs, & Chen, 1990; Jenkins, & Russell, 1960; Palermo, & Jenkins, 1963; Postman, & Keppel, 1970; Van der Made-Van Bekkum, 1971).

Другая вербальная проективная методика — *завершение предложений (sentence completion)* — широко использовалась как в научных исследованиях, так и в клинической практике (P. A. Goldberg, 1965; Haak, 1990; D. H. Hart, 1986; Lah, 1989). С точки зрения длины ответов, степени структурированности и ряда других аспектов, тесты завершения предложений[2] занимают промежуточное положение между методиками словесных ассоциаций и тематическими тестами. Обычно начальные слова или основы предложений допускают почти неограниченное разнообразие возможных завершений, например: «Мое стремление...», «Женщины...», «Меня беспокоит...», «Моя мама...». Основы предложений часто формулируются так, чтобы вызывать ответы, относящиеся к

---

[1] Методика словесных ассоциаций больше не используется таким образом. Обсуждение более современных приложений метода «детектора лжи», или полиграфа, на промышленных предприятиях и в организациях см. в главе 17.

[2] Другое широко используемое название — тесты незаконченных предложений. — *Примеч. науч. ред.*

исследуемой области личности. Подобная гибкость методики завершения предложений представляет одно из ее преимуществ для клинических и научно-исследовательских целей. Тем не менее некоторые стандартизованные формы издавались и для более общего применения.

Примером такой широко используемой формы является Бланк незаконченных предложений Роттера (*Rotter Incomplete Sentences Blank* [*RISB*] — Rotter, & Rafferty, 1950), состоящий из 40 основ предложений. В инструкции тестируемому говорится: «Закончите эти предложения, выражая *ваши искренние мнения.* Попытайтесь проделать это с каждым предложением. Убедитесь, что вы действительно закончили предложение». Каждое завершение оценивается по 7-балльной шкале в соответствии с обнаруженной степенью приспособленности или неприспособленности к действительности. Иллюстративные окончания предложений для каждой оценки даны в руководстве к тесту. Наличие таких образцов ответов позволяет вести достаточно объективный подсчет показателей. Сумма полученных индивидом оценок дает общий показатель приспособленности (*total adjustment score*), который можно использовать в целях скрининга. Содержание ответа можно также проанализировать с клиническими целями для большей спецификации диагностических симптомов. Недавно переработанное руководство по *RISB* включает обновленную нормативную информацию и обзор научных исследований с использованием этого инструмента, проведенных с 1950 г. по настоящее время (Rotter, & Rafferty, 1992).

Множество других тестов завершения предложений было разработано для оценки различных изучаемых популяций и для применения в разнообразных областях научных исследований и психодиагностики[1] (описание некоторых традиционных инструментов см. в работах D. H. Hart, 1986; Lah, 1989; Rabin, & Zlotogorski, 1981). Некоторые интересные современные разработки в этой области включают инструменты, предназначенные для обнаружения симуляции при освидетельствовании на предмет нетрудоспособности, предсказания эффективности управления и оценивания таких конструктов, как защитные механизмы; эти инструменты могут быть полезными и при оценке личности (Carson, & Gilliard, 1993; N. L. Johnson, & Gold, 1995; Timmons, Lanyon, Almer, & Curran, 1993).

Автобиографические воспоминания

Одно из самых последних и многообещающих событий в области проективных вербальных методик — возрождение интереса к использованию *автобиографических воспоминаний (autobiographical memories)* для оценки личности. Анализирование воспоминаний, особенно касающихся ранних этапов жизни, ради понимания возвратных или хронических конфликтов в последующей жизни, разумеется, было главным элементом психодинамической психотерапии со времен Фрейда.[2] Вдобавок, Альфред

---

[1] Одним из широко используемых в научных исследованиях инструментов является Тест завершения предложений Вашингтонского университета (*Washington University Sentence Completion Test* [*WUSCT*]). Разработанный специально для оценки стадий развития эго-концептуализации (*ego-conceptualization*), он рассматривается в главе 16 вместе с другими средствами измерения Я-концепции.

[2] Что касается краткой истории использования автобиографических воспоминаний в оценке личности, см. Bruhn (1995a).

Адлер, один из первых учеников Фрейда, вскоре основавший свою собственную школу индивидуальной психологии, считал, что самые ранние воспоминания, в частности, дают ключ к пониманию «стиля жизни» конкретного человека. В результате этого, психологи-адлерианцы использовали ранние воспоминания как клинические средства и, попутно, в своих исследованиях, начиная с 1930-х гг. (см., например, Hafner, Fakouri, & Labrentz, 1982; Slavik, 1991). Другие теоретики также признавали ту важную роль, которую автобиографические воспоминания, — обычно рассматриваемые как конструкции (= истолкования) или проекции, а не как правдивые исторические отчеты, — могут играть в эволюции личности. В общем, однако, после вспышки интереса в начале XX столетия, к этому, казалось бы, жизненно важному источнику информации о личности не обращались сколько-нибудь систематически вплоть до последнего времени.

С начала 1980-х гг., под влиянием когнитивной точки зрения в психологии, возобновился интерес к автобиографической памяти вообще, и к ее особой функции в организации личности в частности (Bruhn, & Last, 1982; Ross, 1991; Rubin, 1986; Singer, & Salovey, 1993). Выдающийся вклад в эту область исследований связывают с работами Арнольда Брюна (Bruhn, 1984, 1985, 1990a, 1990b). На основе критического анализа моделей, ранее применявшихся фрейдистами, адлерианцами и представителями эго-психологии для интерпретации ранних воспоминаний, он предложил новую концептуальную систему для понимания автобиографической памяти, а также более систематический путь ее использования. В когнитивно-перцептуальной теории (*cognitive-perceptual theory*) Брюна автобиографические воспоминания (*EMs*)[1] являются главными для понимания личности. Поэтому одной из первостепенных задач Брюна стала разработка стандартизованного метода для сбора и интерпретации *EMs*. Процедура ранних воспоминаний (*Early Memories Procedure* [*EMP*] — Bruhn, 1989, 1992a, 1992b) — самоприменяемая бланковая методика для выборочного анализа 21 автобиографического воспоминания, которые относятся ко всему периоду жизни, а не только к детству. Первая часть процедуры предусматривает получение шести общих или «спонтанных» воспоминаний, разграниченных, главным образом, конкретными временными рамками (например, пять самых ранних воспоминаний и воспоминание об особенно важном событии в жизни). Вторая часть процедуры включает 15 конкретных, или «направленных», воспоминаний, зондирующих множество разнообразных событий и областей, которые могут иметь клиническое значение (например, травматическое воспоминание, воспоминание о первом наказании или о самом счастливом моменте в жизни).[2] В добавление к описательным характеристикам каждого воспоминания *EMP* включает разнообразные пробы, касающиеся ясности, эмоционального тона, значимости и многих других признаков воспоминаний. Брюн считает *EMs* специфическими событиями, относящимися к разряду «историй» или метафор, которые отражают то, что люди сознательно усвоили или интуитивно извлекли из своего жизненного опыта. Он также полагает, что эти «истории» часто неточны или искажены,

---

[1] Сокращение *Early Memories* — ранние воспоминания. — *Примеч. науч. ред.*

[2] Инструкции *EMP* также предлагают припомнить «неподобающий сексуальный опыт» и «случаи физического или эмоционального оскорбления», однако респонденты имеют возможность поставить галочки в квадратиках, означающих отсутствие у них подобного опыта. Эта опция, как и бланковая форма *EMP*, снижают возможность получения такого рода «неприличных» воспоминаний о насилии (*abuse*), которые имеют отношение к психическому здоровью как профессионалов, так и широкой публики (см., например, Loftus, 1993).

однако утверждает, что их достоверность несущественна для клинических целей, поскольку, как это бывает с другой проективной продукцией, ценность *EMs* заключена в их способности выявлять имеющиеся заботы, аттитюды, мнения и эмоциональные состояния.

Хотя Брюн и его коллеги разработали (и уже успели переработать) Полную систему количественных оценок ранних воспоминаний (*Comprehensive Early Memories Scoring System* [*CEMSS-R*] — Last, & Bruhn, 1991), подход Брюна к количественному оцениванию и интерпретации автобиографических воспоминаний является довольно гибким.[1] Он смотрит на *EMs* как на сложные психологические феномены, для объяснения которых могут потребоваться различные теоретические модели и, следовательно, разные системы оценивания. Фактически, Брюн пропагандирует разработку заказных систем количественных показателей, или высокоспециализированных систем типа «бутиков», основанных на эмпирически зарегистрированных аспектах *EMs* в критериальных группах и предназначенных для составления научных прогнозов. Брюн и его сотрудники получили многообещающие данные, применяя системы количественных показателей, разработанные для предсказания склонности к делинквентности и насилию (Davidow, & Bruhn, 1990; Tobey, & Bruhn, 1992).

EMP — это методика, которая еще находится в стадии разработки. Работа по сбору норм для EMP даже не начата. И хотя по ряду оценочных категорий, разработанных Брюном и его коллегами, были получены приемлемые уровни согласия между оценщиками, эмпирических данных о других типах надежности *EMP* явно не хватает. К сожалению, получение этих и других психометрических данных в отношении *EMP* может оказаться проблематичным. Как и в случае с другими проективными материалами, сам акт категоризации и квантификации автобиографических воспоминаний неизбежно влечет за собой утрату информации, которая может быть единственно ценной и необходимой для понимания обследуемого человека. Тем не менее *EMP* обладает достаточным потенциалом, чтобы стать весьма полезным инструментом для оценки личности, особенно в контексте психотерапии (см., например, Ritzler, 1993b). Кроме того, систематические выборки автобиографических воспоминаний, вероятно, более показательны в клиническом отношении, чем другие виды вербального материала (рассказанные сновидения, выборки спонтанной речевой продукции и придуманные истории), до сих пор использовавшиеся сходным образом и для тех же целей.[2]

Методики действия

Эта широкая и аморфная категория проективных методик включает довольно много форм относительно свободного самовыражения. Отличительной особенностью всех этих методик является то, что они использовались и как терапевтические, и как диагностические процедуры. Считается, что благодаря предоставляемым ими возможностям выразить себя, человек не только обнаруживает свои эмоциональные за-

[1] Одним из энергично рекомендуемых Брюном технических приемов интерпретации является использование резюме, или кратких сводок, ранних воспоминаний — приема, часто используемого в работе с *TAT* и родственными ему инструментами.

[2] Ряд занятных примеров клинической ценности ранних воспоминаний знаменитых людей, включая сравнение *EMs*, взятых из автобиографических материалов Никсона и Фрейда, см. в Bruhn (1995b).

труднения, но и облегчает свои страдания. К наиболее часто используемым методам из этой категории относятся методики рисования и игровые методики, включая разыгрывание сценок с применением игрушек. Неудивительно, что большинство подобных методов специально разрабатывалось для оценки детей, хотя во многих случаях их можно применять и в работе с взрослыми.

Методики рисования. Несмотря на то что в поисках значимых диагностических признаков для оценки личности исследовались почти все изобразительные средства, техники и темы, особое внимание было уделено рисункам человеческой фигуры.[1] Широко известный пример первых методик этого типа — тест «Нарисуй человека», разработанный Карен Маховер (*Machover Draw-a-Person Test* [*D-A-P*] — Machover, 1949). В этом тесте испытуемый получает карандаш и бумагу с заданием «нарисовать человека». После завершения первого рисунка, его просят нарисовать человека противоположного пола (по отношению к полу только что нарисованного). Пока испытуемый рисует, проводящий тестирование специалист отмечает его реплики, последовательность, в которой рисуются различные части фигуры, и другие подробности процесса рисования. После выполнения рисунков испытуемому обычно задается серия вопросов о возрасте, образовании, профессии и других фактах, имеющих отношение к изображенным персонажам. Это расспрашивание может включать просьбу к испытуемому придумать историю о каждом нарисованном человеке.

Интерпретация *D-A-P*, в том виде как она предполагалась Маховер, является, по существу, качественной и изобилует чрезмерно широкими обобщениями, основанными на взятых по отдельности признаках, например: «Непропорционально большие головы часто рисуются теми, кто страдает органическим заболеванием мозга». Однако данных, оправдывающих подобные утверждения, в руководстве не приводится, даются только ссылки на «тысячи рисунков», проверявшихся в клинических условиях, и для пояснения приводятся несколько выборочных примеров. Систематизированное представление данных в первой публикации теста отсутствует. Вдобавок ко всему, последующие работы по валидизации этого теста, осуществленные другими исследователями, в целом, не смогли подтвердить диагностические интерпретации, предложенные К. Маховер (см., например, Klopfer, & Taulbee, 1976, pp. 558–561).

Другой метод использования рисунков человеческой фигуры (РЧФ), выполненных детьми и младшими подростками, опирающийся на более прочный эмпирический фундамент, был разработан Коппиц (Koppitz, 1968, 1984). Движимая твердым убеждением в клинической полезности РЧФ при обследовании детей, Коппиц разработала и стандартизовала две системы объективных количественных оценок, используя рисунки 1856 учащихся бесплатных школ в возрасте от 5 до 12 лет. Одна из этих систем, основанная, в первую очередь, на тесте рисования Гудинаф—Харриса (см. главу 9) и на клиническом опыте самой Коппиц, используется в качестве возрастного теста умственной зрелости. Вторая система, выведенная из работ Маховер и др., представляет собой проективный тест межличностных аттитюдов и отношений. Она состоит из 30 «эмоциональных индикаторов» (*emotional indicators*), позволяющих отличать рисунки детей с эмоциональными проблемами от рисунков детей, не имеющих подобных проблем. Эти индикаторы (или, попросту, признаки) редко встречались в

[1] Использование рисунков человеческой фигуры в качестве невербального мерила познавательной деятельности рассматривается в главе 9.

рисунках нормальных детей, входящих в выборку стандартизации, и, в отличие от набора индикаторов возрастного развития, предположительно не связаны с возрастом и уровнем зрелости. Они включают: а) *показатели качества (quality signs)*, такие как прозрачности (*transparencies*)[1] и затенение лица; б) *особые признаки (special features)*, такие как крошечные головы или гротескные фигуры, и в) *пропуски (omissions)* некоторых предполагаемых элементов, таких как шея или глаза.

Некоторые совокупности признаков РФЧ, такие как причудливость или суммарное количество «эмоциональных индикаторов», по-видимому, действительно позволяют проводить различие между детьми с проблемами и хорошо приспособленными детьми (D. T. Marsh, Linberg, & Smeltzer, 1991; Naglieri, & Pfeiffer, 1992; Yama, 1990). Однако и сама Коппиц, и другие исследователи предостерегали от использования отдельных индикаторов или «знаков» в диагностических целях. Выработанное общими усилиями мнение по поводу РЧФ состоит в том, что они могут дать лишь самое общее представление об уровне эмоциональной адаптации детей. Более того, когда речь заходит о диагностических приложениях РЧФ, большинство специалистов соглашаются, что рисунки следует использовать только для формулирования гипотез и интерпретировать в контексте другой информации об индивидууме (M. V. Cox, 1993; Knoff, 1993; Tharinger, & Stark, 1990).

Несмотря на эти предосторожности и ограничения, популярность РЧФ не только не ослабла, но, фактически, возросла, свидетельством чему может служить прибавление в семействе заданий на рисование. Одна из наиболее широко используемых методик — «Дом-дерево-человек» (*House-Tree-Person* [*H-T-P*]), которая, как подразумевается ее названием, требует от респондента выполнить отдельные рисунки дома, дерева и человека (Buck, 1948, 1992). Характеристики и особенности самих рисунков, вместе с довольно обширными расспросами после их выполнения, обычно служат источниками гипотез о главных областях конфликта и озабоченности. Более новая методика — Кинетический рисунок семьи (*Kinetic Family Drawing* [*KFD*] — R. C. Burns, 1982; R. C. Burns, & Kaufman, 1970, 1972), — кажется, обладает необычайно высоким потенциалом как клинический инструмент. В этом тесте ребенка просят нарисовать каждого члена его семьи, включая самого себя, «за каким-нибудь занятием». *KFD* породил огромное количество исследований. В недавнем обзоре посвященной ему литературы (Handler, & Habenicht, 1994) утверждается, что, несмотря на методологические проблемы в исследованиях, выполненных с применением этой методики, получен ряд многообещающих результатов, которые оправдывают продолжение исследований при условии использования более тонких методов анализа, таких как множественная регрессия. На данный момент можно предположить, что расстояние и степень взаимодействия между фигурами в *KFD*, например, входят в число наиболее психологически значимых признаков таких рисунков.

Продолжают изобретаться все более творческие задачи на рисование. Например, методика совместного рисования в буквальном смысле требует участия всей семьи или семейной пары для выполнения общими усилиями единственного рисунка в присутствии одного или нескольких терапевтов, внимательно наблюдающих за поведением всех участников (G. Smith, 1991). В основу этой интерактивной методики, ис-

---

[1] Имеется в виду передача освещенности в рисунке. Подробнее об этом см., например: *Арнхейм Р.* Искусство и визуальное восприятие: Пер. с англ. — М.: Прогресс, 1974. — С. 289–293. — *Примеч. науч. ред.*

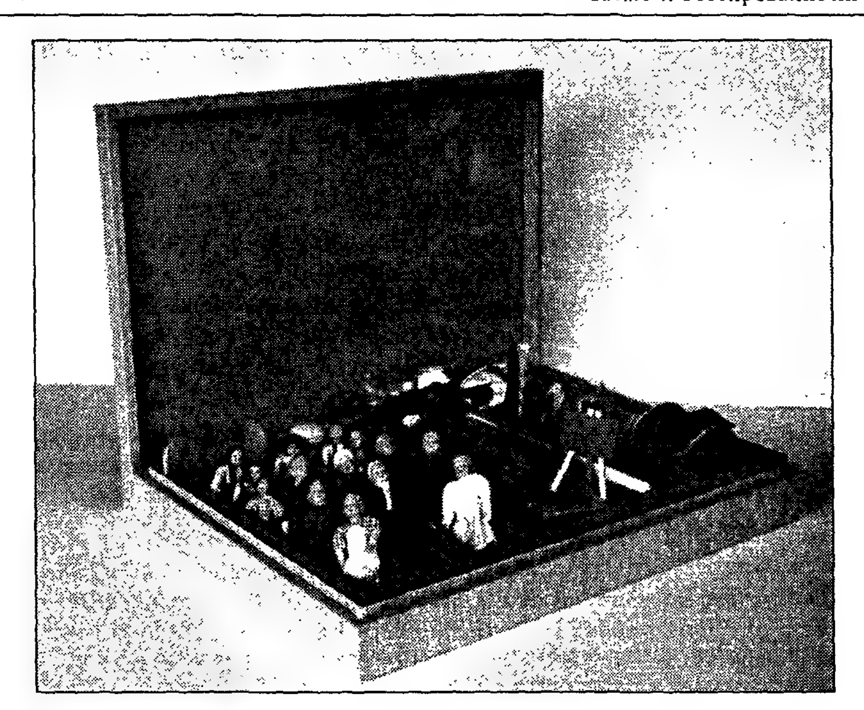


Рис. 15–4. Стандартизованный набор игровых материалов из Сценотеста
(Любезно предоставлен Hogrefe & Huber Publishers. Copyright © by Hogrefe & Huber Publishers. Воспроизводится с разрешения)

пользуемой, главным образом, в контексте семейной терапии, положена идея, позаимствованная из KFD.[1]

Игровые методики и кукольные тесты.[2] Различные виды игровых методик и кукольных тестов, построенные на использовании таких объектов, как марионетки, куклы и игрушечные предметы, получили широкое распространение в проективном тестировании. Ведущие свое происхождение от детской игровой терапии, эти материалы впоследствии были приспособлены для диагностического тестирования взрослых и детей. Объекты отбираются обычно на основе их предполагаемой ассоциативной значимости. Среди наиболее часто употребляемых для этих целей предметов можно назвать кукол, представляющих взрослых и детей обоего пола, игрушечных животных, мебель, кухонные и ванные принадлежности и другие предметы домашнего обихода. Изображенный на рис. 15–4 Сценотест (*The Scenotest*) состоит из стандартизованного набора игровых материалов этого вида в комплекте с руководством. Он был издан в 1960-х гг. в Швейцарии и недавно стал доступен пользователям в США (Staabs, 1991). Предполагается, что игра с подобным материалом обнаруживает отношение ребенка к своей семье, а также соперничество сиблингов, страхи, агрессивность, конфликты и

[1] Подробное обсуждение использования проективных рисунков и критические разборы некоторых из упомянутых в этом разделе методик можно найти в работах Cummings (1986), Hammer (1986), Handler (1996), Knoff (1990).

[2] Что касается общего рассмотрения процедур и интерпретации проективной игры в перспективах психоаналитической теории и возрастного развития, см. Krall (1986).

т. п. Тестирующий отмечает, какие игрушки выбирает ребенок и что он с ними делает, а также его высказывания, выразительные движения и другие факторы внешнего поведения.

С детьми такие методики часто проводятся по принципу свободной игры с имеющимся набором игрушек, поиграть которыми тестирующий предлагает ребенку. При проведении теста с взрослыми игровой материал предлагается с общими инструкциями выполнить определенную, но крайне слабо структурированную задачу. Подобные инструкции могут, конечно, использоваться и при тестировании детей. Зачастую задача имеет признаки сценической постановки, как при расстановке фигурок на миниатюрной театральной сцене. Например, материалы для Сценотеста помещаются в небольшом плоском чемоданчике, крышка которого может использоваться как «сцена» с разнообразными фигурками и декорациями (см. рис. 15–4).

Имеется достаточно полный каталог игровых методик для диагностики и обследования детей (Schaefer, Gitlin, & Sandgrund, 1991). В добавление к проективным инструментам, таким как кукольные методики, в нем представлен широкий ассортимент шкал игровой деятельности (*play scales*) для оценки специфических проблем, от аутизма до гиперактивности, и общих, связанных с возрастным развитием, изменений в таких областях, как мотивация компетентности и темперамент младенца. Включены также шкалы для использования в игровой терапии и для оценивания взаимодействия «родитель—ребенок» и «ребенок—ребенок». По признанию авторов-составителей этого каталога многие из представленных в нем методик находятся еще в начальных стадиях разработки. Тем не менее собранная ими коллекция разнотипных методов включает несколько оригинальных находок и предлагает наряду со строгими высокоструктурированными методами наблюдения, наиболее подходящими для научных исследований, множество клинически ориентированных инструментов.

Оценка проективных методик

Очевидно, что проективные методики весьма заметно отличаются друг от друга. Часть из них выглядит более перспективными по сравнению с другими в силу большей эмпирической поддержки, или большей обоснованности теоретической ориентации, или того и другого вместе. По некоторым методикам, таким как тест Роршаха, были собраны обширные данные, хотя их интерпретация по-прежнему часто вызывает сомнения. О других методиках известно мало либо вследствие их недавнего появления, либо вследствие самой природы таких методик, тормозящей их объективную верификацию, либо вследствие позиции их создателей. Не только в отношении психометрических качеств, но и в отношении существа предъявляемой тестируемому задачи и способов интерпретации результатов, было убедительно доказано, что проективные методики и опросники типа самоотчетов различаются скорее количественно, чем качественно (Levy, 1963).

Отдельные инструменты образуют континуум, на краях которого различия между ними хорошо заметны, тогда как в центральной части многие их характеристики явно перекрываются.

Чтобы оценить каждую проективную методику отдельно и обобщить опубликованную по ней литературу, потребовался бы отдельный том. В рамках этой главы критические замечания вставлялись только в тех случаях, когда дело касалось инстру-

ментов, обладающих уникальными особенностями либо позитивного, либо негативного характера. Можно, однако, поставить несколько общих вопросов, в той или иной степени касающихся большинства проективных методик. Такие вопросы как раз подходят для рассмотрения в форме краткого заключения.

Раппорт и применимость. Большинство проективных методик представляют собой эффективные средства для «растапливания льда» при первых контактах между клиницистом и клиентом. Их задания обычно интересны сами по себе и часто похожи на развлечения. Они ведут к отвлечению внимания индивидуума от самого себя и тем самым к уменьшению смущения и настороженности. И то, что предлагается, почти или совсем не угрожает репутации респондента, так как любой даваемый им ответ является «правильным».

Некоторые проективные методики могут быть особенно полезны при работе с маленькими детьми, неграмотными или с лицами, испытывающими языковые трудности либо страдающими речевыми дефектами. Невербальные средства легко применимы ко всем этим категориям тестируемых. В первых двух группах можно гарантированно получить устные ответы на изобразительные или другие невербальные стимулы. Во всех таких вербально ограниченных группах проективные методики могут помочь проходящему тестирование наладить общение с проводящим его специалистом. Эти методики могут также помочь индивидууму прояснить для себя некоторые стороны собственного поведения, которые до этого оставались невербализованными.

Симуляция. В общем, проективные инструменты в меньшей степени допускают симуляцию, чем опросники типа стандартизованных самоотчетов. Назначение проективных методик обычно замаскировано.[1] Даже если человек имеет некоторый опыт психологической симуляции и знаком с общим смыслом конкретной методики, скажем, теста Роршаха или *TAT*, маловероятно, чтобы он смог предвидеть те сложные способы, которыми будут подсчитываться и интерпретироваться его показатели. Результаты серии недавно проведенных исследований, сравнивающих объективные и проективные средства измерения зависимости, говорят о том, что прослеживается четкая связь между очевидной валидностью теста и степенью, в какой его показатели поддаются фальсификации (Bornstein, Rossner, Hill, & Stepanian, 1994). Кроме того, респондент вскоре полностью погружается в выполнение задания, и потому менее вероятно, что ему удастся прибегнуть к обычным видам маскировки и сохранить сдержанность при межличностном общении.

Вместе с тем было бы неразумно предполагать, что проективные тесты вообще нельзя фальсифицировать. Ряд экспериментов с тестом Роршаха, *TAT* и другими проективными методиками показал, что наблюдаются значительные различия результатов в тех случаях, когда респондентов инструктируют изменить свои ответы с тем, чтобы создать благоприятное или неблагоприятное впечатление, или же когда в инструкциях встречаются формулировки, подтверждающие, что определенные типы ответов более желательны (Masling, 1960). Накоплено большое количество экспериментальных данных в отношении того, что ответы на проективные тесты могут успешно

[1] Степень неведения испытуемого в отношении сущности и цели обследования, в рамках которого эти методики используются, равно как и в отношении замаскированной сущности самих этих методик, представляет собой этическую проблему (см. главу 18).

изменяться как с целью «прикинуться хорошим», так и с целью «показаться плохим», хотя достичь последнего, по-видимому, все же легче. Такие результаты были получены с целым рядом проективных инструментов, включая тест Роршаха, *TAT, P-F Study* Розенцвейга и тесты завершения предложений (Albert, Fox, & Kahn, 1980; Exner, 1991; Kaplan, & Eron, 1965; Meltzoff, 1951; Netter, & Viglione, 1994; Perry, & Kinder, 1990; Schwartz, Cohen, & Pavlik, 1964). Квалифицированный специалист чувствителен к признакам симуляции, обнаруживающимся как в характере отдельных ответов и их паттернах, так и в несовместимости проективных данных с данными о респонденте, полученными из других источников.

Тестирующий и ситуативные переменные. Очевидно, что большинство проективных методик недостаточно стандартизованы в том, что касается проведения или подсчета показателей, либо просто не используются в клинической практике как стандартизованные тесты. Тем не менее имеются доказательства, что даже едва уловимые различия в формулировках словесных инструкций и в отношениях между тестирующим и тестируемым могут заметно изменить результаты этих тестов (Baughman, 1951; Exner, 1993; Hamilton, & Robertson, 1966; Herron, 1964; Klinger, 1966; Klopfer & Taulbee, 1976). Даже когда применяются идентичные инструкции, одни тестирующие в силу своих манер и внешности могут восприниматься ободряющими или успокаивающими, другие — угрожающими. Такие различия могут влиять на продуктивность ответов, выраженность защитной позиции, стереотипию, имажинативность и другие основные характеристики выполнения теста. В свете этих данных проблемы условий организации и проведения тестирования при использовании проективных методик приобретают еще большую важность, чем в случае применения других психологических тестов.

Столь же серьезным минусом оказывается недостаточная объективность процедур подсчета и интерпретации показателей. Даже в тех случаях, когда разработаны и используются объективные системы количественных показателей, конечные шаги в оценке и объединении первичных данных в целостную характеристику обычно зависят от мастерства и клинического опыта специалиста, проводящего обследование с помощью проективных методик. Наиболее беспокоящим следствием такого положения дел является то, что интерпретация показателей часто оказывается столь же проективной для тестирующего, как тестовые стимулы для тестируемого. Другими словами, окончательная интерпретация ответов проективного теста может в большей степени говорить о теоретической ориентации, излюбленных гипотезах и личных особенностях тестирующего, чем о движущих силах личности тестируемого.

Нормы. Еще один бросающийся в глаза недостаток, свойственный многим проективным методикам, имеет отношение к нормативным данным. Такие данные могут или полностью отсутствовать, или быть явно неадекватными, или основываться на нечетко описанных популяциях. При отсутствии адекватных объективных норм клиницист, чтобы проинтерпретировать результаты проективного теста, обращается к своему «широкому клиническому опыту». Но такая система отсчета подвержена всем искажениям памяти, которые сами являются отражениями теоретических предпочтений, предубеждений и прочих индивидуальных особенностей клинициста. Кроме того, контакты любого клинициста могут быть ограничены по большей части людьми, распределения которых по уровню образования, социоэкономическому статусу, полу,

возрасту и другим релевантным характеристикам не являются типичными. По крайней мере, в одном отношении опыт клинициста почти наверняка искажает его представления, так как в силу своей профессии он преимущественно имеет дело с больными или плохо приспособленными к жизни людьми. Поэтому ему может явно не доставать непосредственного знакомства с характерными реакциями нормальных людей на проективные тесты. Нормы по тесту Роршаха, собранные Экснером (Exner), отражают попытки заполнить некоторые из наиболее очевидных пробелов в этой области.

Интерпретация результатов проективных тестов часто связана с подгрупповыми нормами, имеющими субъективную или объективную природу. Такие нормы могут приводить к ошибочной интерпретации, если эти подгруппы не были уравнены в других отношениях. Так, например, если шизофреники и нормальные люди, относительно которых устанавливались нормы, отличались еще и уровнем образования, то замеченные расхождения между выполнением задания шизофрениками и нормальными людьми могут быть следствием разницы в образовании, а не в психическом состоянии. Систематические, или постоянные, ошибки могут, кроме того, давать эффект при сравнении различных психиатрических синдромов. Например, есть некоторые доказательства того, что клиницисты склонны в некоторых этнических группах и у молодых больных вместо диагноза «биполярное расстройство» чаще ставить диагноз «шизофрения»; аналогично этому, не раз сообщалось, что конверсионные расстройства чаще встречаются у лиц с низким социоэкономическим статусом (American Psychiatric Association, 1994).

Надежность.[1] Принимая во внимание особый характер процедур подсчета показателей и недостаточность нормативных данных в проективном тестировании, *надежность оценщика (scorer reliability)* становится важным соображением при оценке методик этого типа.[2] Что касается проективных методик, подходящая мера надежности оценщика должна учитывать не только более объективный, предварительный подсчет показателей, но также завершающие стадии объединения первичных показателей и интерпретации. Недостаточно, например, продемонстрировать, что проводящие тестирование специалисты, овладевшие одной и той же системой определения показателей теста Роршаха, почти сходятся в расчете таких характеристик, как количество ответов на необычные детали, целостных и цветовых ответов. В проективном тесте, подобном тесту Роршаха, первичные количественные меры не могут интерпретироваться непосредственно по таблице норм, как в психологическом тесте традиционного типа. Интерпретационная надежность оценщика касается того, в какой степени различные специалисты по тестированию приписывают одни и те же свойства личности тестируемому на основе своих интерпретаций идентичных протоколов. Этот тип надежности оценщика применительно к проективным тестам практически не исследовался на должном уровне. Некоторые исследователи выявили заметные расхождения в интерпретациях, даваемых достаточно квалифицированными пользователями таких тестов. Принципиальная неоднозначность в таких результатах возникает за счет

---

[1] Содержательное обсуждение вопросов надежности мер тематической апперцепции, с особым акцентом на оценке параметров мотивов, см. в работе С. P. Smith (1992).

[2] Признавая важность обеспечения адекватного уровня объективности подсчета показателей, *Journal of Personality Assessment*, начиная с 1991 г., потребовал от авторов статей, посвященных исследованиям теста Роршаха, предоставлять доказательства, по меньшей мере, 80 % согласия двух или более оценщиков по всем основным категориям показателей.

неизвестного вклада мастерства интерпретатора. Ни высокая, ни низкая надежность оценщика не может непосредственно переноситься на других оценщиков, заметно отличающихся от тех, кто участвовал в конкретном исследовании такой надежности. Фактически, одна из главных причин широкой популярности систем машинной интерпретации для тестов типа Роршаха, как раз и заключается в единообразии их результатов на интерпретационном уровне.

Попытки измерить другие типы надежности тестов в области проективного тестирования оказались еще менее результативными. Коэффициенты *внутренней согласованности (internal consistency)* — в тех случаях, когда они вычислялись, — обычно были низкими. В отношении таких тестов, как тест Роршаха, *TAT* и *P-F study* Розенцвейга, приводились доводы в пользу того, что разные таблицы или задания несравнимы и, следовательно, не должны использоваться при определении надежности методом расщепления теста эквивалентные половины. Фактически, отдельные задания в таких инструментах предназначены для измерения различных переменных. Более того, при интерпретации тенденция ответов на последовательно предлагаемые задания часто рассматривается как значимая переменная. Дж. У. Аткинсон (J. W. Atkinson, 1981; J. W. Atkinson, & Birch, 1978, p. 370–374) с помощью моделирования на ЭВМ доказал, что для процедуры того типа, которая используется в *TAT*, можно добиться высокой конструктной валидности суммарных показателей (например, 0,90) при крайне низкой внутренней согласованности теста (например, 0,07). Он отмечает, что реакции индивидуума на каждую последующую таблицу *TAT* не являются независимыми, но представляют собой непрерывный поток активности, отражающий повышение и снижение относительной силы различных тенденций поведения. Выражение какой-либо из этих тенденций в этой активности ослабляет ее силу. Доля времени, затрачиваемого респондентом на описание, к примеру, мотивированных на достижение действий в ответах на разные таблицы *TAT*, является функцией кумулятивного эффекта реагирования на последовательность таблиц, а также различий стимулов к достижению и других конкурирующих мотивов в ситуациях, представленных на каждой отдельной картинке. С учетом разнообразных аргументов против применения мер внутренней согласованности при оценке надежности проективных тестов, одно из решений проблемы заключается в создании *параллельных форм*, которые были бы действительно сопоставимы, как это сделано в методике чернильных пятен Хольцмана.

Ретестовая надежность также преподносит дополнительные проблемы. При больших временны́х интервалах между сеансами тестирования тест может выявить действительные изменения личности, произошедшие за этот период, при незначительных интервалах повторный тест может оказаться ни чем иным, как припоминанием первоначальных ответов. Когда при повторном проведении *TAT* исследователи давали испытуемым инструкцию сочинить другие истории, для того чтобы определить, будут ли повторяться те же самые темы, большинство учитываемых при подсчете показателей переменных дали незначимые ретестовые корреляции (Lindzey, & Herman, 1955). Здесь также уместно отметить, что многие показатели проективных методик основаны на явно недостаточных выборках ответов. Например, в тесте Роршаха число ответов в рамках протокола обследования конкретного человека, относимых к таким категориям, как движение животного, движение человека, светотени, цвет, необычная деталь и т. п., может оказаться слишком малым, чтобы дать сколько-нибудь надежные индексы. При таких обстоятельствах большие случайные вариации становятся закономерностью, а соотношения и проценты, вычисленные по столь ненадежным мерам, будут даже более нестабильными, чем сами эти меры (Cronbach, 1949, p. 411–412).

Валидность. Для любого теста самым существенным вопросом является вопрос его валидности. Большей частью изучение валидности проективных тестов было сосредоточено на установлении их текущей валидности относительно эмпирических критериев. В основном, в таких исследованиях сравнивалось выполнение теста контрастными группами, скажем, представителями разных профессий или носителями разных психиатрических диагнозов. Однако, как уже отмечалось в связи с обсуждением норм, эти группы часто различаются и в других отношениях, например по возрасту или образованию. В других исследованиях текущей валидности использовался, по существу, метод поиска соответствий (*matching technique*), при котором описания личности, полученные на основании тестовых протоколов, сравниваются с описаниями или данными на тех же людей, взятыми из историй болезни, психиатрических интервью или протоколов длительных наблюдений за поведением. Несколько работ было посвящено изучению прогностической валидности относительно таких критериев, как успехи в специализированных программах обучения, эффективность труда и реакция на психотерапию. Наметилась тенденция к увеличению исследований конструктной валидности проективных методик путем проверки конкретных гипотез, на которых строится использование и интерпретация каждого теста.

Подавляющее большинство опубликованных работ по валидизации проективных методик не позволяют сделать однозначных выводов либо из-за плохой контролируемости условий эксперимента, либо из-за неадекватного статистического анализа, либо из-за того и другого вместе. Некоторые методологические недостатки могут вызывать эффект *мнимого доказательства валидности (spurious evidence of validity)* там, где ее вообще нет, например, при «загрязнении» (*contamination*) критерия или данных теста. Так, эксперты, оценивающие критериальные признаки или критериальную деятельность, могли получить какую-то информацию о выполнении теста конкретным респондентом. Подобным же образом, проводящий тестирование специалист мог получить кое-какие намеки об особенностях респондента либо из разговора с ним во время проведения теста, либо из истории болезни и других, не связанных с тестом, источников. Традиционным средством контроля последнего типа «загрязнения» является «слепой анализ» (*blind analysis*), при котором занесенные в протокол результаты теста интерпретируются специалистом по обработке, не имевшим никаких контактов с респондентом и не имеющим о нем никаких сведений, кроме приводимых в протоколе теста. Однако клиницисты неоднократно пытались доказать, что «слепой анализ» является неестественным способом интерпретации ответов на проективный тест и не согласуется с тем, как проективные инструменты используются в клинической практике.

Другим распространенным источником данных о мнимой валидности является отсутствие кросс-валидизации (Kinslinger, 1966). Из-за большого числа возможных диагностических признаков или засчитываемых элементов, получаемых на основе почти всех проективных тестов, очень легко случайно напасть на множество таких признаков, по которым значимо различаются критериальные группы. Валидность такого определяющего признака может, однако, упасть до нуля, при его применении к новым выборкам.

Иллюстрацией не столь очевидной ошибки служит правильность стереотипа (*stereotype accuracy*). Некоторые описательные формулировки, наподобие встречающихся в протоколе теста Роршаха, можно равным образом применить и к людям в целом, и к молодым юношам, и к госпитализированным больным, и к какой угодно категории испытуемых, выборочно обследуемых с конкретными целями. Соответствие между

критерием и данными теста в том, что касается таких формулировок[1], может создавать ложное впечатление его валидности. Необходимо так или иначе контролировать такие ошибки, например путем измерения соответствия между тестовой оценкой одного респондента и критериальной оценкой другого респондента в той же самой категории. Эта мера указывала бы степень ложного соответствия вследствие правильности стереотипа в условиях конкретного исследования (см., например, L. H. Silverman, 1959).

Еще один распространенный источник ошибок, проистекающих из доверия клиническому опыту при валидизации диагностических признаков, получил название «иллюзорной валидизации» (Chapman, 1967). Этим феноменом можно частично объяснить непрекращающееся клиническое использование диагностических инструментов и систем признаков, эмпирическая проверка валидности которых дала преимущественно отрицательные результаты. В классической серии экспериментов по изучению этого феномена Л. Чепмен и Дж. Чепмен (Chapman, & Chapman, 1967) предъявляли студентам колледжа набор рисунков человеческой фигуры, похожих на рисунки, получаемые в тесте Маховер «Нарисуй человека» (*D-A-P*). Результаты показали, что испытуемые характеризовали эти рисунки с точки зрения сформировавшихся у них расхожих стереотипов, даже если возникающие у них ассоциации не подтверждались данными, с которыми их познакомили во время экспериментального курса «повышения квалификации». Например, они связывали необычные глаза с подозрительностью, большую голову с беспокойством по поводу интеллекта, а широкие плечи с озабоченностью по поводу качеств настоящего мужчины. Эти интерпретации не только оказались не соотнесенными с эмпирическим связями, которые участники эксперимента «изучали на курсах», но, как подтвердили другие эксперименты, такие стереотипные культурные ассоциации почти не поддавались изменению даже в условиях интенсивного обучения, проводимого с целью закрепить противоположные ассоциации. Другими словами, люди оставались верны своим априорным предположениям, даже когда сталкивались с противоречащими им наблюдениями.

Иллюзорная валидизация — конкретный пример механизма, лежащего в основе живучести суеверий. Мы склонны замечать и вспоминать все, что подтверждает наши ожидания, и не замечаем и забываем все, что противоречит им. Подобный механизм может действительно мешать обнаруживать и использовать валидные диагностические признаки тем клиницистам, которые проявляют сильную приверженность к какой-то частной диагностической системе. Оригинальные исследования этой проблемы Чепменами с использованием *D-A-T* были подкреплены аналогичными исследованиями с тестом Роршаха и с Бланком незаконченных предложений (Chapman, & Chapman, 1969; Golding, & Rorer, 1972; Starr, & Katkin, 1969).

С другой стороны, следует отметить, что некоторые недостатки экспериментального плана могут вызвать противоположный эффект, а именно, привести к *недооценке валидности* диагностического инструмента. Общепризнано, например, что традиционные психиатрические категории, такие как шизофрения или синдром деперсонализации, представляют собой грубые классификации тех нарушений, которые в действительности обнаруживаются у больных. Следовательно, если такие диагностиче-

---

[1] Использование таких общеприменимых формулировок есть не что иное, как пример «эффекта Барнума», упоминаемого в главе 17 (Dunnette, 1957; Meehl, 1956). Хорошо сбалансированный обзор обширных исследований этого эффекта см. в Klopfer (1983, p. 510–514).

ские категории используется в качестве единственного критерия для проверки валидности личностного теста, то отрицательные результаты еще не дают основания для однозначных выводов. Аналогично этому, неудача в предсказании критериального признака, связанного с определенной профессией, может отражать лишь факт незнания тестирующим тех черт, которые необходимы для выполнения работы в рамках изучаемой профессии. Когда используются подобные критерии, может случиться так, что данный проективный тест является валидной мерой черт личности, которые он предназначен измерять, но что эти черты не имеют никакого отношения к успеху в выбранных критериальных ситуациях.

Все больше пользователей тестов подчеркивают важность холистического и интегративного принципов в оценке личности, находящих конкретное отражение в учете структуры ответов и контекстуальных переменных при подсчете показателей проективных тестов. Многие из них критиковали непрекращающиеся попытки валидизации отдельных индикаторов, изолированных показателей или диагностических «признаков» (*signs*), получаемых в проективных методиках. Иллюстрацией того, что незначимые корреляции могут оказаться следствием неспособности учесть сложную структуру взаимосвязей личностных переменных, может служить изобилие противоречивых данных о многих проективных методиках, считаемых клиницистами самыми полезными. Например, предполагаемая связь между агрессией в фантазии, выявляемой при помощи *TAT*, и агрессией в реальном поведении не является простой. В зависимости от других, сопутствующих характеристик личности, таких как уровень тревожности или страх перед наказанием, сильная агрессия в фантазии может оказаться связанной как с высоким, так и с низким уровнем физической агрессии (R. Harrison, 1965; Mussen & Naylor, 1954).

Таким образом, отсутствие значимой корреляции между выражением агрессии в историях, сочиненных по картинкам *TAT*, и в реальном поведении лиц, образующих случайную выборку, не является неожиданным, так как эта связь может быть положительной у одних и отрицательной у других. Однако такое отсутствие корреляции, по-видимому, согласуется также и с гипотезой о том, что данный тест вообще не обладает валидностью в отношении выявления агрессивных тенденций. Конечно, в таких случаях необходимы дополнительные исследования, использующие сложные экспериментальные планы, которые допускают анализ условий применимости каждого предположения.

«Проективная гипотеза». Традиционное допущение в отношении проективных методик состояло в том, что ответы индивидуума на предъявляемые ему неоднозначные стимулы отражают существенные и относительно устойчивые свойства личности. Хотя и твердо установлено, что ответы на проективный тест могут отражать и действительно отражают стили реагирования и постоянные черты разных людей, большое и постоянно увеличивающееся число исследований свидетельствует о влиянии на эти ответы множества других факторов. В тех случаях, когда измерялась ретестовая надежность проективных тестов, часто отмечались заметные временны́е сдвиги показателей, указывающие на действие значительной случайной ошибки. Более прямое доказательство чувствительности проективных тестов к временны́м состояниям предоставлено рядом экспериментальных исследований, демонстрирующих влияние на ответы в подобных тестах таких факторов, как голод, недосыпание, допинги, тревога и фрустрация. Были также обнаружены существенные различия в ответах в зависимости от

создаваемых инструкциями установок, характеристик тестирующего и восприятия тестируемым ситуации тестирования. Факторы способности, — и особенно вербальной способности, — несомненно влияют на показатели большинства проективных тестов. В свете всех этих данных понятно, почему ответы в проективном тесте могут обоснованно интерпретироваться только при условии, что тестирующий имеет в своем распоряжении подробную информацию об обстоятельствах, при которых эти ответы получены, а также о способностях и биографии тестируемого.

Преимущества использования неструктурированных, или неоднозначных, стимулов оспаривались представителями других подходов (Epstein, 1966). Такие стимулы столь же неоднозначны для тестирующего, сколь и для тестируемого, а значит, они имеют тенденцию увеличивать степень неопределенности в интерпретации ответов тестируемого. Напротив, при структурированных стимулах имеется возможность отобрать стимулы, релевантные оцениваемым свойствам личности, и менять их характер, чтобы полностью исследовать данное измерение (*dimension*) личности. Такая процедура позволяет дать более ясную интерпретацию результатам теста, чем это оказывается возможным при широкозахватном (*shotgun*) методе неструктурированных стимулов. Существуют также данные, опровергающие распространенное допущение, что чем менее структурированы стимулы, тем с большей вероятностью они будут вызывать проекцию и простукивать «глубинные» слои личности (Klopfer, & Taulbee, 1976; Murstein, 1963). В действительности, эта связь между неоднозначностью и проекцией носит нелинейный характер, с умеренным уровнем неопределенности в качестве оптимума для целей проекции.

Допущение, что фантазия, выявляемая такими проективными методиками, как *TAT*, раскрывает скрытые мотивационные тенденции, также было подвергнуто сомнению. Например, 20-летнее лонгитюдное исследование фантазий в *TAT* и соответствующего реального поведения выявило, что подростковые занятия гораздо лучше предсказывали систему образов, используемых во взрослости при выполнении *TAT*, чем выявляемая с помощью *TAT* образная система подростков — их взрослую деятельность (McClelland, 1966; Skolnick, 1966). Так, участники этого исследования, показавшие восходящую социальную мобильность, став взрослыми, получили более высокие показатели по потребности достижения; но получившие более высокие показатели по потребности достижения в подростковом возрасте не оказались в числе тех, кто впоследствии продемонстрировал восходящую социальную мобильность.

Подобные данные обнаруживают зависимость, прямо противоположную той, что подразумевается в традиционном обосновании проективных методик. Эти данные можно объяснить, если рассматривать ответы в *TAT* не как прямые проективные выражения мотивов, а как выборки мыслей индивидуума, на которые в свою очередь могли повлиять его предыдущие действия и поступки. Те из нас, кто достиг большего, как и те, кто в процессе своего развития чаще сталкивался с ориентированными на достижения моделями поведения, склонны воспринимать в неструктурированных изображениях больше тем, связанных с достижениями.

Подводя итог, можно сказать, что многие типы исследований дали повод для сомнений в самых разных аспектах «проективной гипотезы». Получено достаточно данных, подтверждающих, что ответы человека на неструктурированные, или неоднозначные, тестовые стимулы могут быть объяснены с тем же успехом на основе иных допущений.

Проективные методики как психометрические инструменты. Многие проективные методики явно не выдерживают испытания проверкой на соответствие стандартам тестам. Это очевидно из резюмированных в предыдущих разделах данных, касающихся стандартизации процедур проведения и получения показателей теста, адекватности норм, надежности и валидности. Груда опубликованных исследований, в которых не удалось продемонстрировать хоть какую-то валидность таких методик, как *TAT* и *D-A-P*, поистине впечатляет, даже если сделать скидку на методологические недостатки большинства из них. И все же после нескольких десятилетий преобладания противоречивых результатов использование проективных методик, в сущности, не уменьшилось и, судя по некоторым признакам, может резко возрасти. По словам одного критика, «по-прежнему остаются энтузиасты-клиницисты и сомневающиеся статистики» (Adcock, 1965, р. 533).

Столь явное противоречие можно понять, если признать, что за небольшим исключением проективные методики не являются тестами в подлинном смысле этого слова. К таким широко известным исключениям можно отнести методику чернильных пятен Хольцмана, переработку Экснером методики Роршаха, некоторые адаптации *TAT*, некоторые тесты завершения предложений и *P-Fstudy* Розенцвейга. Разумеется, можно было бы найти и другие, впрочем, немногочисленные примеры квазитестов среди множества проективных методик, не рассматривавшихся в этой главе. Однако даже в отношении этих инструментов нужно накопить значительно больше данных по валидности, чтобы уточнить природу конструктов, измеряемых их показателями, а также собрать гораздо больше нормативных данных на четко определенных популяциях. Таким образом, хотя эти инструменты лучше всех других проективных методик отвечают стандартам тестов, большинство из них не готовы к повседневной эксплуатации для облегчения принятия решений и составления прогнозов в отношении людей.

Проективные методики как клинические инструменты. Вместо того чтобы рассматриваться и оцениваться как психометрические инструменты, или *тесты* в строгом смысле этого термина, большинство проективных методик методики стали восприниматься скорее как клинические инструменты. Так, в руках опытного клинициста они могут служить дополнительным качественным средством интервьюирования клиентов и пациентов. Их ценность как клинических инструментов пропорциональна квалификации клинициста и, следовательно, не может оцениваться независимо от конкретного пользователя. Поэтому попытки оценить их путем применения обычных психометрических процедур, возможно, просто неуместны. И это лишнее доказательство тому, что использование в проективных методиках тщательно разработанных систем получения количественных показателей — не только пустая трата времени, но и источник дезориентации. Подобные методы подсчета показателей придают последним иллюзорную объективность и могут создавать нежелательное впечатление, что данную методику можно рассматривать как тест. Вероятно, особую ценность проективные методики приобретают все же тогда, когда их результаты интерпретируются качественными, клиническими методами, а не в тех случаях, когда результаты их применения обрабатываются количественно и интерпретируются таким образом, как если бы были получены с помощью объективных психометрических инструментов.

Воспользовавшись понятием теории информации, Кронбах и Глезер (Cronbach, & Gleser, 1965) назвали метод интервью и проективные методики «широкополосными» (*wideband*) процедурами. Ширина полосы пропускания или перекрываемого диапа-

зона достигается ценой снижения точности или надежности информации. Объективные психометрические тесты обычно обеспечивают узкий диапазон информации на высоком уровне надежности. В отличие от них, проективные методики и интервью обеспечивают гораздо более широкий диапазон информации, однако менее надежной. Более того, характер полученных посредством одной и той же проективной методики данных может меняться от одного респондента к другому. Например, ответы одного человека по *TAT* могут многое сказать нам о его агрессивности и мало или ничего о его креативности или потребности достижения, и наоборот, ответы другого позволят нам всесторонне оценить степень его креативности и силу потребности достижения, но в них никак не проявится его агрессивность. Подобное отсутствие единообразия в характере информации, получаемой в конкретных случаях, помогает объяснить низкую валидность, обнаруживаемую в тех случаях, когда ответы на проективный тест анализируются относительно какой-либо одной черты на всей группе обследованных лиц.

Интересно отметить, что такая же неровность характеризует интерпретацию клиницистами индивидуальных протоколов обследования. Так, в одном раннем исследовании валидности *TAT* Генри и Фарли (Henry, & Farley, 1959, p. 22) приходят к заключению, что:

> Не существует единственно верного способа применения интерпретации *TAT*. При почти полном отсутствии согласия между экспертами по конкретным пунктам, каждый эксперт принял достаточное количество «правильных» решений, чтобы получился высоко значимый коэффициент согласия. Эксперты могут сделать, в сущности, одни и те же интерпретирующие выводы из протокола теста, но прийти к ним различными путями; или они могут различаться своим умением использовать предсказания *TAT* в различных областях... или в отношении различных тем.

Природа клинической оценки (*clinical judgment*), которая может строиться, помимо прочего, на данных, полученных в результате применения проективных методик и проведения интервью, привлекает все большее внимание психологов (см. главу 17). В этом процессе сами конструкты или категории, исходя из которых организуются данные, постепенно формируются и уточняются индуктивным путем через изучение конкретного сочетания данных, доступных в каждом конкретном случае. Специфическая функция клинициста — делать предсказания на основе уникального или редкого сочетания фактов, для которых невозможно ни подготовить статистические таблицы, ни составить уравнение. Создавая новые конструкты, соответствующие индивидуальному случаю, клиницист может делать предсказания исходя из сочетания фактов, которое ему не приходилось встречать в предыдущих случаях. Делая эти предсказания, он может также принимать во внимание различную значимость сходных фактов для разных лиц. Подобные клинические прогнозы полезны при условии, что они не принимаются как окончательные, а постоянно проверяются относительно информации, извлекаемой из последующего обследования, ответов на тесты, реакций на терапию или другого поведения клиента. Из самой природы интервьюирования и проективных методик следует, что решения не должны основываться на каком-то одном исходном факте или показателе, полученном из таких источников. Эти методики лучше всего подходят для последовательной модели принятия решений, подсказывая в каждом конкретном случае пути дальнейшего исследования или гипотезы для последующей проверки.

16 ПРОЧИЕ МЕТОДИКИ ПСИХОЛОГИЧЕСКОЙ ОЦЕНКИ

Опросники типа стандартизованных самоотчетов и проективные методики, краткий обзор которых дан в предыдущих главах, относятся к числу наиболее известных и широко применяемых инструментов для оценки личности. Тем не менее остается еще богатый запас других методик и приемов, разрабатываемых с той же целью. Это многообразие подходов может привести к появлению методик, которые со временем послужат стимулами прогресса в новых направлениях. Рассматриваемые в этой главе процедуры и приемы являются принципиально исследовательскими методами, хотя некоторые из них могут служить вспомогательными средствами оценки в таких практических областях, как консультирование или организационная психология. Многообразие существующих подходов отображено упоминаемыми здесь конкретными методиками. Некоторые из них довольно трудно классифицировать, поскольку они оценивают конструкты, охватывающие области способностей и личности. Три основные категории включают: а) средства определения когнитивных стилей и типов личности; б) ситуационные тесты и в) методики для оценки Я-концепций и личных конструктов. Чтобы придать более широкую перспективу этому обзору, некоторое внимание будет уделено использованию в оценке личности нетестовых методик, включая натуралистические наблюдения, интервьюирование, рейтинги и анализ биографических данных.

Средства определения стилей и типов

Даже если — в этой ли книге или где-либо еще — тестирование способностей и тестирование личности рассматриваются по отдельности, в реальном использовании тестов и особенно при интерпретации тестовых показателей когнитивную и аффективную сферы просто невозможно разделить.[1] Выборочные образцы поведения, которые в итоге и составляют психологические тесты, представляют собой поперечные

[1] Необходимость интеграции этих областей в общепсихологической теории и исследованиях, равно как и в практической работе, начинает сознаваться все больше. См., например, два недавно вышедших объемистых тома — один под редакцией Саклофски и Зайднера (Saklofski, & Zeidner, 1995), а другой под редакцией Стернберга и Рузгиса (Sternberg, & Ruzgis, 1994), — посвященных вопросам сопряжения (*interface*) интеллекта и личности.

срезы поведенческого репертуара какого-то конкретного человека и как таковые они одновременно содержат информацию обо всех аспектах этого человека. В главе 11, например, мы рассматривали доказательства растущего согласия по поводу того, что способности не могут исследоваться независимо от аффективных переменных, потому что выполнение тестов способностей явно зависит от стремления к достижениям, настойчивости, ценностей и т. п. Точно так же в главе 13 мы обсуждали взгляд, согласно которому люди различаются по межситуационной устойчивости поведения в зависимости от того, как они воспринимают и категоризуют ситуации, что, в свою очередь, зависит от их прошлого опыта и научения. Более того, во всех главах 3-й и 4-й частей этой книги мы рассматривали измерительные средства, якобы нацеленные на оценку либо способностей, либо личности, которые тем не менее учитывают аспекты, относящиеся к аффективному или когнитивному функционированию соответственно.

Справиться с многообразием факторов, свойственным человеческому поведению, помогает целый ряд способов. Мы можем, к примеру, просто изучать корреляцию между мерами разных черт или свойств, таких как тревожность и способность решать задачи (см., например, Zeidner, 1995). Можно также воспользоваться методами многомерной статистики, такими как факторный анализ и многомерное шкалирование, чтобы выделить и классифицировать компоненты в границах множества поведенческих данных (Jones, & Sabers, 1992). Можно к тому же разрабатывать структурные схемы, которые определяют места для нескольких переменных и топографически отображают взаимосвязи между ними. Разработанная Верноном иерархическая модель структуры способностей (см. главу 11), шестиугольное представление профессиональных тем Холланда (см. главу 14) и межличностная круговая модель (см., например, Hofstee et al., 1992) — все это примеры структурной организации в рамках отдельных областей поведения.[1] Еще один подход состоит в том, чтобы использовать категории, которые сами по себе достаточно сложны и охватывают как когнитивные, так и аффективные элементы. Эмоциональный интеллект (Salovey, & Mayer, 1990; Mayer, & Salovey, 1993) — один из таких конструктов, который недавно привлек к себе повышенное внимание (см., например, Goleman, 1995). Когнитивные стили и типы личности, обсуждаемые в двух ближайших разделах, также служат примером такого подхода. Они представляют собой попытки зафиксировать качественные различия в структуре или форме человеческого поведения.

Когнитивные стили. *Когнитивные стили (cognitive styles)* указывают, по существу, на предпочитаемые конкретным человеком и типичные для него способы восприятия, запоминания, мышления и решения задач (Messick et al., 1976). Они рассматриваются как широкие стилистические особенности поведения, которые являются сквозными характеристиками способностей и личности и проявляются во многих видах деятельности и способах действия. По различным когнитивным стилям и близким к ним понятиям стилей научения (*learning styles*) и стилей мышления (*thinking styles*) накопилась обширная исследовательская литература (Brodzinski, 1982; Furnham, 1995; Globerson, & Zelniker, 1989; Goldstein, & Blackman, 1978a, 1978b; Grigorenko, & Sternberg, 1995; Jonassen, & Grabowski, 1993; Kogan, 1976; Messer, 1976; Sternberg, 1994; Witkin, & Goodenough, 1981).

[1] Упоминавшееся в главе 14 сферическое представление профессиональных интересов, предложенное Трейси и Раундсом (Tracey, & Rounds, 1996), являет собой пример более сложной, многомерной схемы.

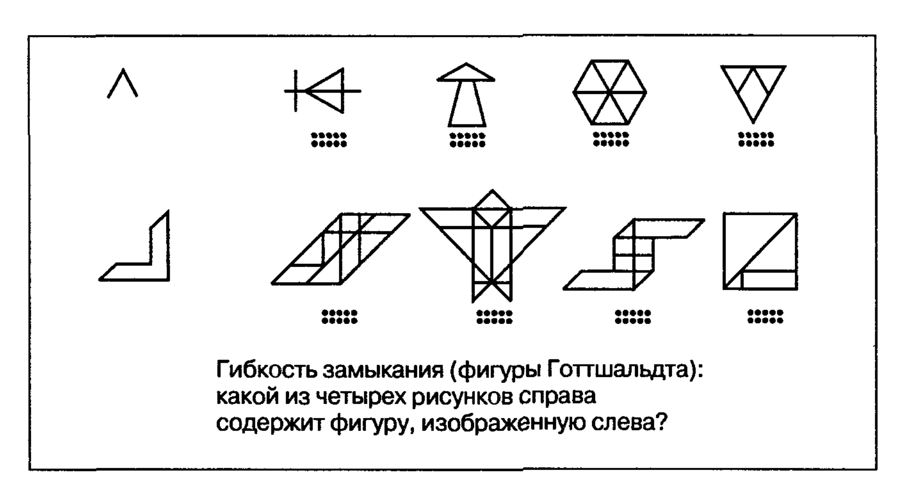


Рис. 16–1. Задания, иллюстрирующие тип перцептивных задач, используемых при оценке личности. *(Из Thurstone, 1950, р. 7)*

Один из основных источников дифференциации когнитивных стилей можно обнаружить в области перцептивных функций. В многочисленных публикациях по результатам экспериментальных исследований в этой области убедительно показаны значимые связи между характерными особенностями аттитюдов, мотивационной или эмоциональной сферы индивидуума и выполнением им перцептивных или когнитивных заданий. Следует также признать, что ряд проективных методик, в особенности тест Роршаха, являются, по существу, перцептивными тестами (см., например, Blatt, 1990).

Среди факторов, выявленных в ранних факторно-аналитических исследованиях восприятия, особенно плодотворным для исследования личности оказался один — гибкость замыкания (*flexibility of closure*)[1] (Pemberton, 1952; Thurstone, 1944). Распространенный тип теста для оценки этого фактора требует от испытуемого распознавания геометрической фигуры среди отвлекающих и запутывающих деталей. Два задания из такого теста, имеющего высокую нагрузку по этому фактору (фигуры Готтшальдта), показаны на рис. 16–1. В нескольких ранних исследованиях были получены требующие теоретического осмысления данные, указывающие на возможные связи между этим перцептивным фактором и чертами личности. В одном из них, например, лица, получившие высокий показатель по гибкости замыкания, имели также высокие самооценки по таким описательным характеристикам черт, как социальная скромность, независимость от мнения других, склонность к анализу, интерес к теоретическим и научным проблемам, а также нелюбовь к наведению порядка и рутине (Pemberton, 1952). В последующие годы разнообразные адаптации фигур Готтшальдта широко использовались в исследованиях когнитивного и некогнитивного поведения.

Подходя к этой проблеме с других позиций, Г. Виткин и его сотрудники (Witkin et al., 1954/1972) в ходе длительных исследований воспринимаемой ориентации в пространстве выявили такую важную переменную, как способность сопротивляться разрушающему влиянию конфликтующих контекстуальных признаков-подсказок (*cues*). С помощью разнообразных тестов, в которых использовались стержень и рамка, дви-

[1] Имеется в виду гибкость замыкания фигуры, или завершения гештальта. — *Примеч. науч. ред.*

гавшиеся независимо друг от друга, а также кресло и комната, менявшие угол наклона, эти исследователи смогли показать, что люди сильно различаются по своей «полезависимости» (*field dependence*), т. е. по тому, в какой степени на восприятие ими вертикали влияет окружающее зрительное поле. Показатели двух типов надежности тестов, определявшейся путем корреляции четных и нечетных замеров и повторного тестирования, были высокими, а большинство корреляций между различными тестами пространственной ориентации оказались значимыми. Таким образом, был собран большой объем данных, свидетельствующих, что полезависимость является относительно устойчивой, постоянной чертой, обладающей определенной степенью обобщенности.

Еще больший интерес представляют значимые корреляции между этими тестами ориентации и Тестом встроенных фигур (*Embedded Figures Test* [*EFT*])[1], который считался средством измерения полезависимости в чисто зрительной ситуации бланкового теста (весьма похожего на фигуры Готтшальдта, приведенные на рис. 16–1). Когда накопилось больше эмпирических данных, измерение «полезависимость–поленезависимость» стали рассматривать как перцептивную составляющую более широкого измерения личности, названного когнитивным стилем «глобальность—расчлененность», или психологической дифференциацией (Witkin, Dyke, Faterson, Goodenough, & Karp, 1962/1974). Существуют эмпирические доказательства, что этот когнитивный стиль обнаруживает большую устойчивость на протяжении детства и ранней взрослости и связан с рядом личностных переменных, таких как лидерство (Weissenberg, & Gruenfeld, 1966), социальная конформность (Witkin et al., 1974) и др. (см., например, Jonassen & Grabowski, 1993, chap. 7).

Масштаб и разнообразие исследований полезависимости поистине впечатляющие, от межличностных отношений (Witkin, & Goodenough, 1977) до научения и памяти (D. R. Goodenough, 1976), успехов в усвоении математики (Vaidya, & Chansky, 1980), выбора области обучения в колледже и аспирантуре (Raskin, 1985), межкультурных различий (J. W. Berry, 1976) и предпочтений в отношении рабочей среды (Wooten, Barner, & Silver, 1994). Пример интригующих связей, которые выявляются в обзорах многочисленных исследований, относится к выводу о том, что поленезависимые склонны придерживаться активных, «участвующих» подходов к учению, тогда как полезависимые чаще пользуются «зрительскими» подходами. Одно недавнее исследование, в котором результаты тестов с множественным выбором ответов сравнивались с оценками на основе анализа выполнения учебных заданий (*performance-based assessments*), позволяет заключить, что последние благоприятствуют поленезависимым учащимся (Lu, & Suen, 1995). С другой стороны, в межличностных ситуациях полезависимые, вероятно, имеют некоторые преимущества в том, что касается умения ладить с другими. Они склонны быть более внимательными к социальным сигналам, более чуткими к поведению других и более эмоционально открытыми, чем поленезависимые. По-видимому, ни один из концов континуума «полезависимость—поленезависимость» не является непременно или единственно благоприятным либо неблагоприятным; скорее всего, ценность отклонений в любом направлении зависит от требований конкретных ситуаций.

В большинстве этих исследований использовался Тест встроенных фигур, который относительно прост в проведении. Издаются формы этого теста для детей (от-

---

[1] В отечественной литературе название этого теста часто переводят как «тест замаскированных фигур» (см., например, *Бурлачук Л. Ф., Морозов С. М.* Словарь-справочник по психодиагностике. — СПб.: Питер Ком, 1999. — С. 111–112). — *Примеч. науч. ред.*

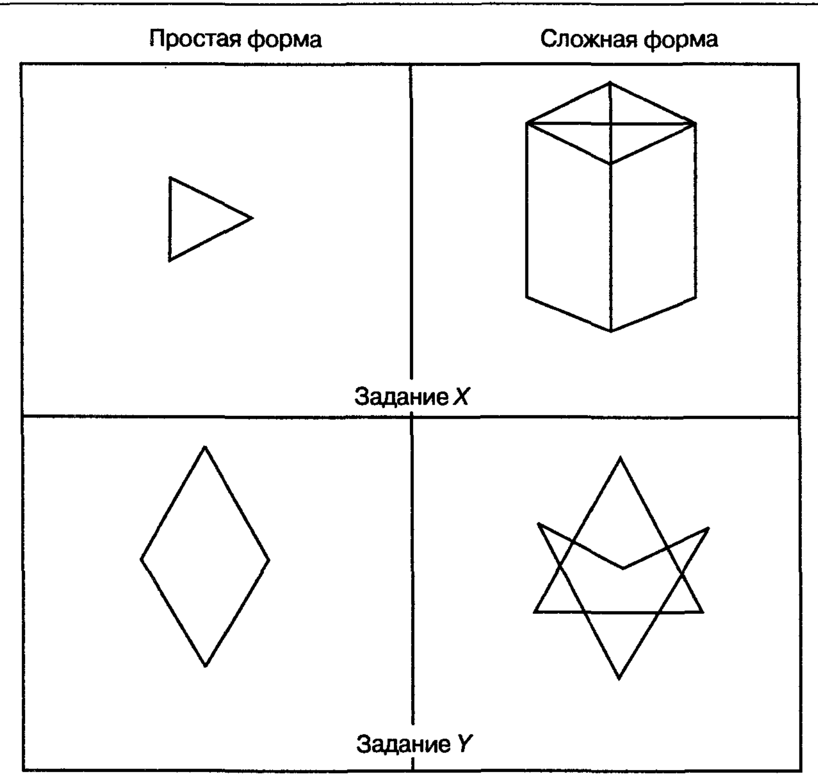


Рис. 16–2. Демонстрационные задания из Группового теста встроенных фигур *(Copyright © 1972 by Consulting Psychologists Press. Воспроизводится с разрешения)*

дельно — для дошкольников) и взрослых, а также для группового проведения (Coates, 1972; Witkin, Oltman, Raskin, & Karp, 1971). Два демонстрационных задания из Группового теста встроенных фигур (*Group Embedded Figures Test*) приведены на рис. 16–2. В формах для детей дошкольного и более старшего возраста сложные фигуры представляют собой относительно легко узнаваемые, знакомые по опыту объекты, а сами эти тесты проводятся индивидуально. Исходная взрослая форма *EFT* также рассчитана на индивидуальное проведение.

Хотя поток исследований по этой теме никогда не иссякал (см., например, Bertini, Pizzamiglio, & Wapner, 1985), многое еще предстоит сделать, чтобы разобраться в противоречивых результатах, не так уж редко встречающихся в исследованиях поленезависимости и ее связей с широким множеством поведенческих переменных. Одна проблема касается терминологии и, в свою очередь, отражает разногласия по поводу того, что такое поленезависимость. Одни исследователи пришли к выводу, основываясь на корреляциях этой характеристики с мерами фактора *g*, что конструкт «поленезависи-

мость» означает все-таки когнитивную способность; другие же отнесли поленезависимость к категории когнитивного «контроля» (*control*), занимающей промежуточное положение между когнитивными способностями и когнитивными стилями (McKenna, 1984; Jonassen, & Grabowski, 1993). Аналогичное препятствие для обобщений в этой области представляет неоднородность состава участников исследований и отсутствие единообразия в методологии. Например, полученные данные дают возможность предположить, что, тогда как на результаты измерений поленезависимости с помощью бланковых методик (типа Группового теста встроенных фигур) сильное влияние оказывает общий интеллект, показатели тестов действия, наподобие Портативного теста стрежня и рамки (*Portable Rod-and-Frame*), являются более чистым отражением поленезависимости как свободной от значений стилистической переменной (Arthur, & Day, 1991).

Несмотря на концептуальные и методологические помехи, присутствующие в исследованиях когнитивных стилей, сами эти стили представляют большой теоретический и практический интерес.[1] Поскольку они находятся на линии пересечения способностей и личности, эти стили могут оказывать модулирующее воздействие на поведение как в эмоциональной, так и в интеллектуальной области. Поэтому вопросы об их природе и характеристиках, — наподобие вопроса о степени постоянства или, наоборот, гибкости когнитивных стилей, — приобретают большую важность (см., например, Niaz, 1987). Более того, там, где научение представляет собой главную цель, все больше признается, что эффективность обучения и оценки его результатов можно существенно повысить за счет учета стилистических факторов (Furnham, 1995; Jonassen, & Grabowski, 1993; Lu, & Suen, 1995; Sternberg, 1988, 1994b; Zelniker, 1989).

Типы личности. Подобно когнитивным стилям, *типы личности (personality types)* тоже относятся к конструктам, которые использовались для объяснения внутри- и межиндивидных сходств и различий в предпочитаемых способах мышления, восприятия и поведения. По существу дела, типы личности — это категории, определяемые конфигурациями двух или большего числа черт или атрибутов. В качестве средств объяснения человеческого поведения, типологии имеют длинную историю, начало которой восходит к древнегреческой теории «соков тела» — крови, черной желчи, желтой желчи и слизи — и связываемым с ними темпераментами: сангвиническим, меланхолическим, холерическим и флегматическим. Типологические системы часто обладают огромной притягательностью для непрофессионалов, поскольку предлагают относительно простые и, на первый взгляд, твердые принципы понимания и, возможно даже, объяснения собственного поведения и поведения других людей. Такие системы сильно различаются по числу и виду измерений (*dimensions*), которые могут использоваться для порождения типов.

В рамках психологии за годы ее существования как науки был разработан целый ряд различных типологий. Одни из них, подобно системе Шелдона для классификации типов личности на основе телосложения (Sheldon, & Stevens, 1942/1970), осно-

[1] Следует отметить, что когнитивные стили различаются по сложности. Некоторые стили касаются относительно простых различий, — например, такой стиль, как «импульсивность–рефлексивность», или когнитивный темп (*cognitive tempo*), определяется только скоростью, с какой люди реагируют на проблемные ситуации, особенно неопределенные (Kagan, 1965; S. B. Messer, 1976). Другие стили, наподобие предложенных Стернбергом стилей мышления (Sternberg, 1988, 1994b), охватывают более сложные структуры предрасположенностей.

вывались, главным образом, на умозрительных построениях и были со временем отброшены как не нашедшие эмпирического подтверждения. Другие, подобно описанным в главе 13 типам кодов *MMPI*, выросли из эмпирических наблюдений, получили развитие и продолжают использоваться до сих пор (Graham, 1993). Большинство типологических систем охватывают различия внутри какой-то одной области, например профессий или темпераментов, и лишь немногие имеют дело более чем с одной областью. К числу последних можно отнести попытки Алана Миллера (Alan Miller) синтезировать существующие типологии по всем выделенным когнитивным, аффективным и мотивационным измерениям (*dimensions*). Предложенные Миллером (Miller, 1991a, 1991b) концептуальные связующие элементы основаны на результатах аналитического обзора значительного числа типологий в этих трех областях и, судя по всему, должны представлять большую эвристическую ценность.

Индикатор типов Майерс—Бриггс. Одна из самых живучих типологических классификаций была разработана К.-Г. Юнгом (1921/1971), и она же послужила основанием для создания Индикатора типов Майерс—Бриггс (*Myers—Briggs Type Indicator* [*MBTI*] — Myers, 1962; Myers, & McCaulley, 1985) — инструмента, нашедшего широкое применение в области оценки личности нормальных людей.[1] В *MBTI* используется известная юнгианская дихотомия экстравертированной и интровертированной установок (*E* и *I*), а также его классификация противоположных способов чувствования или, употребляя более современную терминологию, способов получения информации (ощущение или интуиция — *S vs. N*), и полярных подходов к оценке (мышление или чувство — *T vs. F*). Кроме того, в нем учитывается предпочтение одной из двух ориентаций в отношении внешнего мира (суждение или перцепция — *J vs. P*)[2] — полярность, не представленная эксплицитно в работе Юнга. Показатели по этим четырем, предположительно независимым, измерениям (*dimensions*) дают в результате 16 возможных «формул типов», которые задаются комбинациями букв, обозначающих предпочитаемое направление в каждом из четырех измерений. Например, комбинация *INTP* означает «интровертированный интуитивно-мыслительный и перцептивно-ориентированный» тип (*introverted, intuitive, thinking, and perceptive type*).[3] Каждое предпочитаемое направление получает также балльную оценку, которая пока-

---

[1] *MBTI* кратко упоминался в главе 13 в связи с использованием метода вынужденного выбора в личностных опросниках типа стандартизованных самоотчетов.

[2] С этим измерением (*judgment vs. perception*) из-за неудачно выбранных названий полюсов связано много недоразумений даже среди англоязычных пользователей опросника, не говоря уже о тех, кто работает с его переводами. У нас часто переводят эти названия как «рациональный—иррациональный» (см., например, *Бурлачук Л. Ф., Морозов С. М.* Словарь-справочник по психодиагностике. — СПб.: Питер Ком, 1999. — С. 167), что неверно и с лингвистической, и с психологической точки зрения. Здесь нет возможности рассматривать все встречающиеся интерпретации измерения *J vs. P*, но если учесть контекст обсуждения MBTI, связанный с рассмотрением стилистических характеристик способностей и личности, одна из таких интерпретаций выходит на первый план. Люди с высокими *J*-показателями тяготеют к завершенности (*closure*), не терпят незаконченных дел, половинчатых решений и испытывают чувство облегчения и удовлетворенность после достижения определенности. Люди с высокими *P*-показателями, напротив, противятся быстрому замыканию гештальта, оттягивают принятие окончательных решений, стремясь сохранить возможность выбора (*open-ended*) и накопить побольше данных, а сделав окончательный выбор, нередко испытывают тревогу и жалеют о своей поспешности. — *Примеч. науч. ред.*

[3] Согласно руководству к *MBTI*, такой человек, помимо всего прочего, был бы: а) тихим и скрытным, б) получающим удовольствие от решения проблем с опорой на логику и анализ, в) интересующимся преимущественно идеями и г) имеющим вполне определенные интересы (Myers, & McCaulley, 1985).

зывает силу предпочтения и подсчитывается на основе разности баллов между полюсами каждого измерения.

Результаты *MBTI*, в отличие от большинства других личностных опросников, предназначены главным образом для использования самими респондентами и потому представляются в безоценочной форме. Две самые важные предпосылки, на которых основана интерпретация результатов *MBTI*, состоят в следующем: а) все типы личности одинаково ценны, необходимы и имеют свои сильные и слабые стороны; б) люди всегда более компетентны и искусны в предпочитаемых ими функциях, процессах и установках. Эти отличительные особенности способствовали популярности *MBTI* и его применению для самых разнообразных целей, включая профориентацию, консультирование и комплектование рабочих групп. Психометрические проблемы, поставленные показателями *MBTI*, хорошо документированы (что касается критических разборов, см. DeVito, 1985; J. S. Wiggins, 1989). В частности, критики выдвигали возражения против вынужденных дихотомий *MBTI*, из-за чего идентичные комбинации букв, обозначающих тип, присваиваются на основе разности показателей, которые могут существенно варьировать по абсолютной величине. Тем не менее с этим инструментом продолжает проводиться значительное количество исследований. В добавление к тем, что нацелены на изучение валидности *MBTI*, часть исследований была посвящена проверке и сравнению альтернативных способов подсчета его показателей (см., например, Davis, Grove, & Knowles, 1990; Girelli, & Stake, 1993; Harvey, & Murry, 1994). Кроме того, были разработаны варианты *MBTI* и адаптации его конструктов для различных направлений использования этого опросника. Один из заслуживающих упоминания вариантов — недавно разработанный Вопросник для оценки стилей учащихся (*Student Styles Questionnaire*), предназначенный для оценки стилей учения, практической работы и рассказывания (*relating*) у учащихся в возрасте от 8 до 17 лет. Несмотря на популярность Индикатора типов Майерс—Бриггс и еще нескольких типологий, история сложилась так, что большинство психологов, посвятивших себя изучению личности, понимали ее скорее как некий психический составной объект (*psychic entity*), собранный из компонентов, по которым люди различаются количественно, чем как основанную на качественных различиях таксономию. К тому же в сфере научной психологии имело место общее сопротивление использованию типов как объяснительных категорий. Это неприятие типов проистекало, в основном, из трех источников: а) подчеркивания многими психологами важности понимания и оценки самобытности каждого человека; б) имплицитной связи между типами и стереотипами и связанных с последними опасностей; в) нехватки адекватных количественных методов для установления достоверности и анализа категорийных данных.

В серии последних работ, однако, некоторым из традиционных возражений против типологических конструктов был брошен вызов. Например, Пол Мил (Paul Meehl) считает, что на смену периоду пользования произвольно выбранными, искусственными категориями пришло время эмпирического поиска и исследования возможного «залегания» естественных классов, или таксонов, в непатологической области индивидуальных различий личности. Он и его коллеги предложили новые таксонометрические (*taxometric*) методы, которые можно использовать для определения таких классов (Meehl, 1992, 1995; Meehl, & Golden, 1982; Meehl, & Yonce, 1994).[1] Подобным

[1] Несмотря на практическое отсутствие доказательств переносимости из одной ситуации в другую, подгруппы, получаемые в результате кластерного анализа биографических данных (рассматриваемых несколько позже в этой главе), также предлагают интересные — с точки зрения типологической классификации — возможности.

же образом, после критического анализа целого ряда геометрических моделей черт личности, от одномерных биполярных черт до многомерных комплексов, Далстром (Dahlstrom, 1995) делает вывод, что такие модели неадекватны задаче организации конструктов личности и ее исследования. Согласно Далстрому, принципиально целостная (*configural*) природа личности и столь же целостное ее функционирование требует «ящичной классификационной схемы, чтобы в строгой и точной манере резюмировать все то, чем мы отличаемся и чем похожи друг на друга» (р. 14). Наконец, даже глубоко укоренившееся мнение о том, что стереотипы обычно неточны и вредны, недавно было подвергнуто сомнению (Lee, Jussim, & McCauley, 1995).

Эти и другие события предвещают возобновление интереса и появление более тонкого подхода к использованию многомерных категорий, в частности типов и стилей, как части концептуального арсенала, пригодного для изучения и объяснения индивидуальных различий в поведении. Тем не менее, какими бы полезными или популярными ни оказались такие конструкты, они будут, без всякого сомнения, существовать не вместо, а вместе с подходами, подчеркивающими важность неповторимых особенностей конкретных людей и их поведения. При психологической оценке людей всегда нужно помнить об опасностях материализации типов либо их упрощенческого или негибкого использования в качестве объяснительных понятий.

Ситуационные тесты

Хотя термин «ситуационный тест» получил распространение во время и сразу после Второй мировой войны, соответствующие этому термину тесты были разработаны несколько раньше. По существу, ситуационным называется тест, при котором тестируемый помещается в ситуацию, моделирующую или очень похожую на критериальную ситуацию из «реальной жизни». Таким образом, эти тесты обнаруживают некоторое принципиальное сходство с методиками получения выборочных образцов работы (*job-sample techniques*), используемых при конструировании тестов профессиональных достижений, и с методиками оценки учебных достижений на основе анализа реальной деятельности (*performance-based assessments*).[1] Однако, критериальное поведение, выборочные замеры которого производятся в ситуационных тестах, обычно бывает более изменчивым и сложным. Кроме того, выполнение ситуационных тестов оценивается не на основе способностей и знаний, а в «единицах» таких личностных переменных, как эмоциональные реакции, межличностные отношения и аттитюды.

Тесты для программы «Исследование воспитания характера». Среди самых первых ситуационных тестов — хотя в то время они так не назывались, — были тесты, разработанные Х. Хартшорном (Н. Hartshorne), М. Мэем (М. А. Мау) и их сотрудниками (1928, 1929, 1930) для программы «Исследование воспитания характера» (*Character Education Inquiry* [*CEI*]). Эти тесты предназначались, главным образом, в качестве исследовательских инструментов для использования в обширном проекте по изучению природы и развития характера у детей. Тем не менее эти методы можно приспособить для других целей тестирования, и примеры такого применения уже имеются.

---

[1] Характеристика этих двух классов методик дана в главе 17.

В общем, в методиках *CEI* использовались знакомые, естественные ситуации из повседневной жизни школьников. Тесты проводились в виде плановых контрольных работ в классе, как часть домашнего задания ученикам, в ходе спортивных соревнований или в групповых играх. Кроме того, дети не сознавали, что их тестируют, кроме случаев, когда в процедуру тестирования включался обычный школьный опрос. В то же время все тесты Хартшорна—Мэя представляли собой тщательно стандартизованные инструменты, дающие объективные количественные показатели.

Тесты *CEI* предназначались для измерения новым и оригинальным способом таких поведенческих характеристик, как честность, самоконтроль и альтруизм. Большинство из них касались честности и были связаны с ситуациями, в которых детей заставляли поверить в то, что они могут схитрить или обмануть, не боясь быть разоблаченными. Например, в задаче с кругами (*Circles Puzzle*) ребенка просили, закрыв глаза, поставить метки в каждом из 10 маленьких случайно расположенных кругов. Контрольное тестирование в исключавших подглядывание условиях показало, что получить в сумме трех попыток результат свыше 13 правильных помет практически невозможно. Поэтому показатель свыше 13 регистрировался как доказательство подглядывания и, следовательно, нечестности.

Большинство тестов *CEI*, как оказалось, обладают хорошей различительной способностью, обеспечивая широкий диапазон индивидуальных различий в показателях. Надежность их также, по-видимому, вполне удовлетворительна. Однако ответы детей обнаружили существенную ситуационную специфичность. Интеркорреляции различных тестов, принадлежащих к одной категории (например, тесты на честность или упорство), оказались очень низкими. Эта специфичность становится вполне понятной, когда мы рассматриваем действие интересов, ценностей и мотивов конкретного ребенка в различных ситуациях. Например, ребенок, у которого сильно выражен мотив выделиться в учебе, не обязательно заинтересован в достижениях на спортивных соревнованиях или в групповых играх. Эти различия в мотивации могли, в свою очередь, сказаться на поведении ребенка в тестах на честность, проводимых в этих различных контекстах. Как в отношении полученных данных, так и в их интерпретации, исследователи, участвовавшие в программе *CEI*, предугадали на четыре десятилетия раньше то значение специфичности, которое придается этому понятию в наше время. Однако проведенный позднее повторный анализ данных *CEI*, позволил предположить, что более подходящая для их объяснения модель включает в себя как общий фактор честности, так и ситуационную компоненту, и что есть даже некоторые основания для выделения общего фактора «морального характера» (Burton, 1963; Rushton, 1984).[1]

Ситуационные тесты в центрах оценки и методики разыгрывания ролей. Ситуационные тесты составляли большую часть программы центров оценки, которую Бюро стратегических служб США (OSS) ввело в период Второй мировой войны. Методика оценки в центрах (*assessment-center technique*) требует, по существу, пребывания в стационаре в течение нескольких дней, когда кандидатов наблюдают и испытывают в различных ситуациях и разными способами. Эта методика представляла собой главную процедуру при отборе кандидатов для службы в военной разведке (Murray, & MacKinnon, 1946; OSS Assessment Staff, 1948). Аналогичные процедуры были впослед-

---

[1] Интересный анализ роли, которую исследования Хартшорна и Мэя сыграли в полемике по поводу привития добродетелей и формирования характера, можно найти в работе Vitz (1990).

ствии внедрены в Институте оценки и исследования личности при Калифорнийском университете, а также включены в ряд крупномасштабных проектов оценки военнослужащих и гражданских специалистов.[1]

Примером одного типа тестов, разработанных *OSS*, может служить ситуационный тест «Напряженная ситуация», предназначенный для выборочного изучения поведения человека в условиях стресса, фрустрации или эмоционального срыва. Например, перед обследуемым ставится задача, которую нужно выполнить с двумя «помощниками», которые, в действительности, не только не помогают, но и мешают ее выполнению. В ситуационном тесте другого типа использовалась *группа без лидера (leaderless group)* в качестве средства для оценивания таких характеристик, как умение работать в команде, находчивость, инициатива и лидерство. Предлагаемая в таких тестах задача требует совместных усилий группы испытуемых, причем никто из них не назначается главным и не наделяется особыми полномочиями. Примеры из программы *OSS* включают ситуацию «Водная преграда» (*Brook Situation*), требующую переброски людей и оборудования через небольшую речку с максимальной скоростью и безопасностью, и ситуацию «Преодоление укрепления», в которой людей и снаряжение нужно было переправить через две «крепостные стены», разделенные воображаемым рвом.

Вариантом этой методики является Групповая дискуссия без лидера (*Leaderless Group Discussion* [*LGD*]). Не требующая для своего проведения много времени и особого оборудования, методика *LGD* широко применялась при отборе таких групп, как армейские офицеры, инспектора и управляющие государственных служб, руководители среднего звена и кандидаты на высшие управленческие должности в промышленности, торговые агенты и руководители отделов сбыта, учителя и социальные работники. По существу дела, группе просто дается тема для обсуждения в течение установленного времени. Проводящие тест специалисты наблюдают и оценивают каждого участника дискуссии по целому ряду критериев, но не участвуют в самой дискуссии. Хотя методика *LGD* часто использовалась в неконтролируемых и нестандартизованных условиях, она стала предметом значительного числа психометрических исследований. Их результаты свидетельствуют о том, что особенно в тех случаях, когда оценщики должным образом подготовлены, *LGD* может быть эффективным инструментом предсказания эффективности труда, предъявляющего повышенные требования к вербальному общению и решению проблем в ходе дискуссии, а также к умению завоевывать признание среди сослуживцев (Bass, 1954; Greenwood, & McNamara, 1967; Guilford, 1959; Thornton, & Zorich, 1980).

Некоторые ситуационные тесты используют *разыгрывание ролей (roleplaying)* или импровизацию, для того чтобы вызвать интересующее поведение. Фактически, упоминавшиеся выше групповые дискуссии без лидера, да и некоторые другие ситуационные тесты, можно рассматривать как простые или усложненные варианты разыгрывания ролей. Хотя разыгрывание ролей было одним из методов, используемых в программе оценки *OSS*, оно имеет более ранние истоки и более широкое применение. Исчерпывающий обзор истории, теоретических основ и разнообразных вариантов метода разыгрывания ролей или импровизации можно найти в работе McReynolds, & DeVoge (1978). Этот метод предполагает, что обследуемому человеку дается прямая инструкция сыграть некую роль либо открыто (с партнерами или без них), либо в

---

[1] Описание этих узловых проектов оценки персонала, начиная с программы *OSS*, можно найти в книге J. S. Wiggins (1973/1988, chap. 11).

форме рассказа о том, что бы он в этом случае делал и/или говорил. Сама ситуация может быть представлена реалистически, как на театральной сцене, либо задана в форме аудио-, видеозаписи или печатного текста.[1]

Метод импровизации продолжает пользоваться огромной популярностью, хотя и используется по большей части нестрого и всякий раз приспосабливается к конкретной области и локальным условиям. Одно важное приложение этого метода — профессиональная оценка персонала, особенно когда межличностное поведение занимает важное место в должностных обязанностях (Stricker, 1982; Stricker, & Rock, 1990). Особый пример — оценка эффективности консультанта. В этой ситуации за будущим консультантом либо непосредственно наблюдают во время проводимого им сеанса консультирования «подготовленного клиента» (*coached client*) — штатного сотрудника или сокурсника в роли клиента, излагающего заранее отобранные и стандартизованные проблемы, либо анализируют видеозапись такого сеанса (см., например,Connor, 1994, p. 72–75; Kelz, 1966; Neufeldt, Iversen, & Juntunen, 1995; A. Williams, 1995). Когда используется видеозапись, проходящий обучение может пронаблюдать свои действия и оценить их, в дополнения к оценкам экспертов и сокурсников.[2]

Методики оценки в центрах неоднократно доказывали свою эффективность в роли предикторов разнообразных критериев (см., например, Coulton, & Feild, 1995; Howard, & Bray, 1988; Ritchie, 1994; Tziner, Ronen, & Hacohen, 1993). Они широко использовались в областях, где критерии отбора являются сложными, например, в области профессий, связанных с правовым принуждением (J. L. Coleman, 1987; Moore, & Unsinger, 1987). Была даже проведена некоторая работа по адаптации этих методик для оценки глухих претендентов на получение работы (Berkay, 1993). Вследствие широкой применимости таких методик мы не в состоянии сделать какие-то обобщения по поводу валидности любой *конкретной* методики оценки в центре, — результаты различаются от центра к центру, в зависимости от используемых в них специфических процедур, характера критерия и квалификации экспертов. В общем, коэффициенты валидности обычно оказываются самыми высокими там, где самые надежные методологии — такие, как в программах обследования, использующих несколько способов оценки, включая оценки лиц одной категории с испытуемым, — и где акцент делается на релевантных и доступных прямому наблюдению измерениях (*dimensions*) поведения (Gaugler, Rosenthal, Thornton, & Bentson, 1987; Shore, Shore, & Thornton, 1992; Thornton, & Byham, 1982). Несмотря на обширные исследования, ряд вопросов в отношении методик оценки в центрах остаются без ответа. Вероятно, самый неприятный вопрос касается невозможности в нескольких исследованиях продемонстрировать конвергентную валидность различных методов оценивания отдельных измерений (*dimensions*) выполнения тестовых заданий.[3]

---

[1] Мультимедийные и интерактивные компьютерные технологии делают возможным применение совершенно новых способов предъявления реалистичных стимульных ситуаций и вариантов реакций на них со стороны тестируемого. Обсуждение возможностей и проблем в разработке мультимедийных тестов и описание инструментов с использованием интерактивного видео, измеряющих навыки разрешения конфликтов, можно найти в работах Drasgow, Olson-Buchanan, & Moberg (1996) и Olson-Buchanan, Drasgow, Moberg, Mead, & Keenan (1996).

[2] В клинической психологии разыгрывание ролей подверглось всесторонней и систематической разработке в рамках нескольких теоретических подходов, и особенно в программах модификации поведения, семейной терапии и супружеского консультирования.

[3] Краткие, но информативные критические обзоры исследований в центрах оценки см. в Landy, Shankster, Kohler (1994, p. 277–278); Schmidt, Ones, Hunter (1992, 635–637).

Представления о себе и личные конструкты

Восьмидесятые и девяностые годы свидетельствовали о возрождении интереса к Я-концепции и связанным с ней конструктам (Byrne, 1996; Harter, 1990; Hattie, 1992; Markus, & Wurt, 1987; Oosterwegel, & Oppenheimer, 1993; Wylie, 1989).[1] Ряд современных подходов к оценке личности сосредоточены на том, как люди смотрят на себя и других. Такие методы часто отражают влияние феноменологической психологии, в которой основное внимание уделяется тому, как события *воспринимаются* конкретным человеком. Тем самым описание себя (или самохарактеристика) по праву приобретает первостепенную важность вместо того, чтобы рассматриваться в качестве второстепенных заместителей других поведенческих наблюдений. Внимание исследователей сосредоточивается также на степени самопринятия, проявляемого индивидуумом.

Все рассматриваемые в этом разделе методики нацелены на оценку результатов восприятия человеком себя и других людей. Хотя лишь несколько таких инструментов распространяются специальными издательствами, большинство методик разрабатывалось для конкретных исследовательских проектов и их можно легко найти в соответствующих публикациях. Одни из них представляют интерес, главным образом, вследствие связи с определенными теориями личности или с областью активных, непрекращающихся исследований. Другие находят широкое применение при изучении самых разных проблем.

Тест завершения предложений Вашингтонского университета. Вероятно, можно утверждать, что тесты для измерения Я-концепции по существу не отличаются от опросников типа самоотчетов, обсуждавшихся в главе 13. И все-таки точнее было бы говорить, что стандартизованные самоотчеты фактически являются средством измерения Я-концепции. Интерпретация ответов на опросники с точки зрения самоконцептуализации (*self-conceptualization*) составляет основу теоретического подхода к развитию личности, сформулированного Дж. Ловингер (Loevinger, 1966a, 1966b, 1976, 1987, 1993; Loevinger, & Ossorio, 1958). Пытаясь свести воедино множество разнородных данных из своих собственных и других исследований, Дж. Ловингер высказала предположение о существовании свойства личности, которое она определила как способность концептулизировать себя, или «дистанцироваться» (*assume distance*) от себя и своих импульсов. Согласно Дж. Ловингер, проявления именно этого свойства в ответах на личностные опросники были описаны в таких терминах, как эффект фасада (*facade*), оборонительная позиция тестируемого (*test-taking defensiveness*), установка на ответ (*response set*), социальная желательность (*social desirability*), молчаливое согласие (*acquiescence*) и личный стиль ответов (*personal style*). Подобно некоторым другим психологам, Дж. Ловингер рассматривает такие аттитюды тестируемого не как инструментальные ошибки, которые должны быть исключены, а как главный источник значимой дисперсии в личностных опросниках.

На основе данных из многочисленных источников Дж. Ловингер высказала предположение, что способность формировать Я-концепцию или, иначе говоря, составлять представление о себе, увеличивается с возрастом, уровнем развития интеллекта, образованием и повышением социоэкономического статуса. В самой низкой точке

[1] Обзор философских и психологических подходов к «Я» (*the self*) и связанным с ним процессам, охватывающий период от XVII по XX в. включительно, можно найти в книге Levin (1992).

развития, примером которой служит младенец, индивидуум не способен к самоконцептуализации. По мере развития этой способности ребенок постепенно формирует стереотипизированное, конвенциональное и социально приемлемое представление о себе. Эту стадию Дж. Ловингер считает типично подростковой. По мере продвижения к зрелости индивидуум преодолевает такие стереотипные представления в направлении дифференцированной и реалистичной Я-концепции. На этой стадии люди полностью сознают свои отличительные особенности и признают себя такими, какие они есть в действительности.

Именно такую черту личности как способность к самоконцептуализации, получившую название «развитие эго», или «уровень эго», Дж. Ловингер с сотрудниками и попытались измерить с помощью Теста завершения предложений Вашингтонского университета (*Washington University Sentence Completion Test* [*WUSCT*] — Loevinger, 1985, 1987; Loevinger, & Wessler, 1970; Loevinger, Wessler, & Redmore, 1970). В соответствии с теоретической позицией его авторов постулируется девять уровней развития эго: досоциальный (*Presocial*), импульсивный (*Impulsive*), защиты собственных интересов (*Self-Protective*), конформиста (*Conformist*), самоанализа (*Self-Aware*), сознательности (*Conscientious*), индивидуалистический (*Individualistic*), автономный (*Autonomous*) и интегрированный (*Integrated*). Кроме первого уровня, предшествующего появлению вербальных навыков, остальные уровни можно оценить с помощью *WUSCT*. Каждое завершенное предложение получает ранговую оценку уровня эго, на основе которых рассчитывается совокупный показатель по полному тесту. *WUSCT* основан на исследованиях, проводившихся с женщинами и девочками-подростками, и только позже был адаптирован для применения к мужчинам и мальчикам. Еще позже тест был переработан и теперь пользователям его доступны сравнимые формы для мужчин и женщин (Loevinger, 1985; Novy, 1992). В новом руководстве по подсчету показателей *WUSCT* представлены данные как для женской, так и для мужской выборок (Hy, & Loevinger, 1996).

Дальнейшие исследования с применением *WUSCT* подтвердили его надежность (Novy, & Francis, 1992; Weiss, Zillberg, & Genevro, 1989) и валидность как средства измерения конструкта «развитие эго» в различных выборках (Bushe, & Gibbs, 1990; Novy, Gaa, Frankiewicz, Liberman, & Amerikaner, 1992; Westenberg, & Block, 1993). Особенно плодотворным оказалось применение этого инструмента при изучении половых различий в развитии эго (Cohn, 1991). Один из немногих проблематичных аспектов *WUSCT*, разделяемый с другими тестами завершения предложений и методиками, допускающими свободные вербальные ответы, заключается в том, что его показатели имеют тенденцию коррелировать с беглостью речи и словарным запасом, а в некоторых случаях могут подвергаться прямому влиянию этих переменных (см., например, Vaillant, & McCullough, 1987; Westenberg, & Block, 1993). Хотя эту возможность нужно учитывать при планировании исследований с *WUSCT*, равно как и при использовании этого теста с другими целями, она согласуется с его теоретическим обоснованием и не должна делать недействительными получаемые результаты.

Инвентари самооценок и родственные измерительные средства. Во все большей массе исследований конструкт «Я-концепция» сливается с довольно родственными конструктами, обозначаемыми как самооценка (Baumeister, 1993; Bedner, & Peterson, 1995) и воспринимаемая самоэффективность (Bandura, 1982, 1995; Maddux, 1995; Schwarzer, 1992). Главной опорой этих исследований является воздействие самооцен-

ки индивидуума на эффективность его деятельности. В действительности, самооценка (*self-esteem*) типично характеризуется как оценочный компонент Я-концепции. На основе длительного, кумулятивного воздействия такие самооценки могут влиять на развитие когнитивных и аффективных черт. В частности, имеет место широкое согласие по поводу того, что самооценка является определяющим фактором таких психологически важных переменных, как совладающая способность (*coping ability*) и чувство благополучия (*sense of well-being*).

Конструкт «самооценка», на первый взгляд, обманчиво прост. Часто предполагается существование общего оценочного отношения к себе, варьирующего от крайне положительного до крайне отрицательного, которое является устойчивым и всецело субъективным по характеру. Измерение самооценки в исследовательских и практических целях традиционно проводилось исходя из этих предположений. И эта традиция стала настолько прочной, что Бласкович и Томака (Blascovich, & Tomaka, 1991) в своем исчерпывающем критическом обзоре средств измерения самооценки и Я-концепции отнесли разработанную Розенбергом (Rosenberg, 1965) Шкалу самооценки (*Self-Esteem Scale* [*SES*]) — десятипунктную, прямую (*face valid*) шкалу типа самоотчета — к наиболее часто используемым средствам такого рода.

Простые и очевидные меры общей самооценки, наподобие *SES*, кажется, и в самом деле относительно устойчивы во времени и, в ряде случаев, могут быть полезными в качестве способа оценки общего самоуважения (*self-regard*). Тем не менее, многие исследователи пришли к выводу, что отношение между Я-концепцией и поведением можно продемонстрировать с большей ясностью, если первую рассматривать как иерархический, многомерный конструкт — и оценивать соответственно (Fleming, & Courtney, 1984; Marsh, Byrne, & Shavelson, 1992; Marsh, & Shavelson, 1985; Shavelson, & Bolus, 1982; Shavelson, Hubner, & Stanton, 1976; Uguroglu, & Walberg, 1979). При определенных обстоятельствах единая, глобальная мера самооценки может дать противоречивые результаты или не обнаружить значимых корреляций с другими переменными, в то время как более узко определяемый конструкт, такой как академическая Я-концепция, даст согласующиеся и значимые результаты. Это особенно справедливо в отношении тех, для кого учебные достижения могут не занимать высокого места в личной системе ценностей. В таких случаях общее высокое мнение о себе, необъективно отражающее собственную систему ценностей респондента, может не коррелировать значимо с учебными достижениями или интеллектуальной деятельностью. Кроме того, исследования показывают, что меры «переживаемой самооценки», основанные на самоотчетах (*experienced self-esteem*), и меры «проявляемой самооценки» (*presented self-esteem*), основанной на отчетах других, не всегда коррелируют (Demo, 1985).

Подобные результаты и выводы привели к тому, что в последние годы исследователи отошли от одномерных концептуализаций самооценки и переключились на изучение ее специфических аспектов, и эта тенденция быстро набирает силу. Особенно много исследований, посвященных взаимосвязи академической Я-концепции и академических достижений у детей и подростков. В этой конкретной области, исследования, использующие сложную методологию типа моделирования структурными уравнениями и лонгитюдных планов, позволяющую анализировать направление причинных связей, подтвердили точку зрения, согласно которой академические Я-концепции связаны с конкретными учебными предметами. Они также показали, что эти Я-концепции коррелируют, предсказывают и влияют на последующие академические достижения (Fortier, Vallerand, & Guay, 1995; House, 1995; Lyon, & MacDonald, 1990; H. W. Marsh, 1990a, 1990b).

При создании новейших средств измерения Я-концепции была использована обширная теоретическая и эмпирическая литература, накопившаяся в этой области. Шкала Я-концепции школьника (*Student Self-Concept Scale* [*SSCS*] — Gresham, Elliott, & Evans-Fernandez, 1993), например, является серийно выпускаемым и доступным для приобретения инструментом, при создании которого авторы использовали в качестве отправной точки разработанную А. Бандурой (Bandura, 1982, 1986) теорию самоэффективности, опираясь также на другие теории и результаты исследований. *SSCS* оценивает три главные области Я-концепции, а именно, академическую, социальную и образ себя (*self-image*). Внутри каждой из этих областей респонденты указывают не только степень своей уверенности в том, что они способны сделать из перечисленного в пунктах шкалы, но и степень значимости для них каждого из пунктов, а также степень уверенности в том, что, обладая определенными качествами или сделав определенные вещи, они достигнут определенных результатов. *SCSS* предусматривает получение показателей по подшкалам и совокупного показателя относительно раздельных статистических норм для учащихся мужского и женского пола на каждом из уровней начальной и средней школы.

Аналогичный, хотя и не выпускаемый для свободной продажи, измерительный инструмент — Инвентарь личной и академической Я-концепции (*Personal and Academic Self-Concept Inventory* [*PASCI*] — Fleming, & Whalen, 1990) — разрабатывался для оценки учащихся средней школы и студентов колледжей. Он отражает попытки авторов операционализировать и более глубоко исследовать иерархическую, многоаспектную модель Я-концепции, развитую Шейвлсоном (Shavelson) и его коллегами. *PASCI* представляет собой четвертую редакцию и расширение экспериментальной шкалы, разработанной еще в 1950-х гг. для оценки чувства неадекватности или несостоятельности (*feelings of inadequacy*).[1] Современная версия включает шкалу общей самооценки и шесть дополнительных шкал частных аспектов Я-концепции. Из этих шести две шкалы имеют дело с социальными аспектами Я-концепции (Социальное принятие и Социальная тревожность), еще две имеют отношение к ее физическим аспектам (Внешность и Физические способности), а две оставшиеся являются академическими (Способности к математике и Вербальные способности). Разумеется, что и *SSCS*, и *PASCI* подвержены систематическим ошибкам в ответах, присущим всем стандартизованным самоотчетам и рассмотренным в главе 13. Кроме того, оба этих инструмента еще нуждаются в дополнительном документальном подтверждении их эффективности в тех контекстах, для которых они разрабатывались. Однако, с точки зрения дифференциации конструкта и определения содержания, они достаточно полно отражают достижения в концептуализации и измерении самооценки.

Контрольный список прилагательных. Несколько широко ориентированных методик было разработано специально для оценки представлений человека о самом себе. Одна из них — Контрольный список прилагательных (*Adjective Check List* [*ACL*]) — пользуется немалой популярностью и доступна для приобретения. Первоначально сконструированная для использования в научно-исследовательской программе Ин-

[1] Бласкович и Томака (Blascovich, & Tomaka, 1991) дают краткий обзор эволюции этого инструмента, также как и многих других мер самооценки. Модель Шейвлсона была одной из наиболее влиятельных в этой области и использовалась также в концептуализации *SSCS*. Более подробную информацию об этой модели можно найти в Marsh et al. (1992), Marsh & Shavelson (1985), Shavelson, & Bolus (1982), Shavelson et al. (1976).

ститута оценки и исследования личности (*IPAR*), эта методика содержит список из 300 расположенных в алфавитном порядке прилагательных от *absentminded* до *zany*[1] (Cough, 1960; Cough, & Heilbrun, 1983). Респонденты отмечают все прилагательные, которые они считают своими описательными характеристиками.

В своем современном виде ACL предусматривает получение показателей по 37 шкалам, из которых четыре шкалы оценивают установки на ответ. Шкалы, образующие главный кластер, первоначально составлялись с опорой на целесообразность или содержание, путем подбора прилагательных для каждой из 15 потребностей, входящих в перечень Мюррея и охватываемых EPSS (см. гл. 15). Дополнительный набор из 9 «актуальных шкал» (*topical scales*) разрабатывался преимущественно путем привязки пунктов к эмпирическому критерию, с тем чтобы получить меры черт, считающихся важными в межличностном поведении. Два оставшихся кластера шкал проектировались с опорой на специализированные теории личности: транзактный анализ Э. Берна (Berne, 1961, 1966) и теорию креативности и интеллекта Уэлша (Welsh, 1975b). Главным источником данных эмпирической валидизации всех 37 шкал, к которому обращались на том или ином этапе разработки каждой шкалы, было непосредственное наблюдение участников программ оценки в центрах *IPAR* и итоговые рейтинги черт. На основе этих оценок, проведенных в *IPAR*, и данных других, дополнительных исследований в руководстве к *ACL* приводятся характеристики личности, соответствующие высоким и низким показателям по каждой шкале.

В качестве исследовательского инструмента *ACL* применялся к ошеломляющему многообразию проблем, относящихся к таким областям, как психопатология, выбор профессии, творчество, политическая и экономическая деятельность и даже реакции пациентов на устройства для исправления неправильного прикуса и контактные линзы. *ACL* также использовался для оценки личности исторических персонажей на основе их биографий и публикаций (Welsh, 1975a) и для характеристики неодушевленных объектов, таких как города или автомобили. Позднее эта методика применялась, среди прочих, при изучении возрастных изменений у женщин в середине жизни (Helson, & Wink, 1992; Wink, & Helson, 1993; York, & John, 1992) и в исследованиях нарциссизма (Wink, 1991, 1992).[2]

Q-сортировка. Еще одной специальной методикой, пригодной для исследования *Я*-концепции, является *Q*-сортировка, первоначально разработанная Стефенсоном (Stephenson, 1953) в целях инструментального обеспечения подхода, известного как *Q*-методология (см., например, Kerlinger, 1986, chap. 32; McKeown, & Thomas, 1988). В этой методике респонденту дается набор карточек, содержащих какие-либо утверждения или названия черт, которые он должен рассортировать на «кучки» — от «наиболее характерных» до «наименее характерных» для него. Формулировки утверждений или названия черт могут браться из стандартного перечня, но чаще для большего соответствия конкретному случаю их подбирают специально. Чтобы обеспечить равнораспределенность оценок (*ratings*), используется «принудительно-нормальное» (*forced normal*) распределение, когда респонденту дается инструкция класть в каждую «кучку» строго определенное количество карточек. Такой расклад можно составить

---

[1] «Рассеянный» и «сумасбродный» соответственно. — *Примеч. науч. ред.*

[2] Что касается рецензий на *ACL*, см. Teeter (1985) и Zarske (1985). Полную библиографию по ACL до 1980 г. включительно можно получить у издателей (Gough & Heilbrun, 1980).

для выборки названий (или формулировок) любого объема, обратившись к таблице нормального распределения. Следует отметить, что, подобно рассмотренной в главе 13 методике вынужденного выбора, *Q*-сортировка дает ипсативные, а не нормативные данные. Иными словами, респонденты сообщают нам о том, что они считают своими сильными и слабыми качествами, но не о том, насколько сильно, по их мнению, эти качества выражены у них по сравнению с другими людьми или какой-либо внешней нормой.

Q-сортировка использовалась при изучении разнообразных психологических проблем (Bem, & Funder, 1978; Block, 1961/1978; Kogan, & Block, 1991; Ozer, 1993; Rogers, & Dymond, 1954). При тщательном исследовании личности конкретного человека его часто просят провести повторную сортировку того же набора карточек, но в рамках иной системы отсчета. Например, респондент может сортировать карточки применительно к себе самому и к другим людям, скажем своему отцу, матери, жене или мужу. Точно так же он может проводить сортировки применительно к самому себе, но в разных ситуациях, скажем на работе, дома или при взаимодействии с другими людьми. *Q*-сортировку можно также применять для получения от респондентов сведений о том, какими, по их мнению, они является на самом деле (реальное «Я»), какими их видят другие люди (социальное «Я») и какими они хотели бы быть (идеальное «Я»). Для наблюдения за изменениями, происходящими у индивидуума, *Q*-сортировки можно получать последовательно на разных стадиях психотерапии — процедура, которой чаще других придерживались приверженцы клиенто-центрированной терапии. В процессе терапии представление о себе меняется в сторону более благоприятных оценок и приближается к идеальному «Я» индивидуума (Rogers, & Dymond, 1954, chap. 4).[1]

Семантический дифференциал. Эта методика впервые была разработана Ч. Осгудом и его сотрудниками (Osgood, Suci, & Tannenbaum, 1957) как инструмент для исследований по психологии значения, хотя его возможности для оценки личности были быстро осознаны. Семантический дифференциал представляет собой стандартизованную, дискретную процедуру для измерения коннотативных значений любого данного концепта у конкретного человека. Каждый концепт оценивается по 7-балльной графической шкале как более тесно связанный с одним или другим из пары полюсов шкалы (рис. 16–2). Для каждого концепта применяется серия — обычно из 15 или более — биполярных шкал, заданных словами, употребляемыми в качестве прилагательных. Корреляционный и факторный анализ первоначального набора из 50 разработанных Осгудом шкал выявил три основных фактора: 1) фактор *оценки (Evaluative)*, с высокими нагрузками по таким шкалам, как хороший—плохой, полезный—бесполезный, чистый—грязный; 2) фактор *силы (Potency)*, обнаруженный в таких шкалах, как сильный—слабый, большой—маленький, тяжелый—легкий; 3) фактор *активности (Activity)*, идентифицированный в таких шкалах, как активный—пассивный, быстрый—медленный, острый—тупой. Фактор «оценки» является самым весомым, объясняющим наибольший процент полной дисперсии.

---

[1] Все эти процедуры можно, разумеется, реализовать с ранее описанным Контрольным списком прилагательных *ACL* или любыми другими методиками для оценки Я-концепции. Вдобавок ко всему, Q-сортировки и контрольные списки могут использоваться для протоколирования оценок наблюдателей, что обсуждается в этой главе позднее.

ОТЕЦ

Хороший	: : : : : : : : :	**Плохой**
Чистый	: : : : : : : : :	**Грязный**
Жестокий	: : : : : : : : :	**Добрый**
Медленный	: : : : : : : : :	**Быстрый**
Полезный	: : : : : : : : :	**Бесполезный**
Напряженный	: : : : : : : :	**Расслабленный**
Сильный	: : : : : : : : :	**Слабый**
Большой	: : : : : : : : :	**Маленький**

Рис. 16-3. Иллюстрация методики семантического дифференциала. Оценивая концепт «отец», респондент ставит галочку в соответствующем сегменте каждой шкалы. Обычно число используемых шкал значительно больше

Ответы по семантическому дифференциалу можно анализировать несколькими способами. При количественной обработке оценкам по каждой шкале могут приписываться числовые значения от 1 до 7 или от –3 до +3. Тогда можно измерить общее сходство любых двух концептов для отдельного человека или группы исходя из их положения на всех шкалах. Коннотативные значения всех концептов, оцененных одним человеком, можно проанализировать посредством вычисления «показателя» каждого концепта по трем описанным выше главным факторам. Так, на шкале с числовыми значениями от –3 до +3 концепт «мой брат» у конкретного человека может получить –2 по фактору оценки, 0,1 по фактору силы и 2,7 по фактору активности.

Оцениваемые концепты могут выбираться в соответствии с любой изучаемой проблемой. Респондентов можно, например, попросить оценить: а) самих себя, членов своих семей, друзей, коллег по работе, учителей или общественных деятелей; б) членов различных этнических или культурных групп; в) представителей различных профессий; г) разнообразные занятия, такие как учеба или спортивные игры на открытом воздухе; д) абстрактные понятия, такие как смущение, ненависть, болезнь, мир или любовь; е) названия товаров или торговые марки; ж) радио- или телепрограммы. Семантический дифференциал применялся в самых различных контекстах, в исследованиях по таким разным проблемам, как клиническая диагностика и терапия, выбор профессии, культурные различия и реакции потребителей на товары и торговые марки (Snider, & Osgood, 1969). Кроме того, не прекращается работа по совершенствованию самой этой методики (см., например, Cogliser, & Schriesheim, 1994). Библиография по семантическому дифференциалу насчитывает более 2000 источников.

Репертуарный тест ролевых конструктов. Одной из методик, специально созданных для использования в клинической практике, является разработанный Дж. А. Келли (G. A. Kelly, 1955, 1963, 1970) Репертуарный тест ролевых конструктов или, сокращенно, Реп-тест (*Role Construct Repertory Test* [*Rep Test*]). Разработка Реп-теста тесно связана с теорией личности Дж. Келли. Основная идея, на которой строится эта теория, состоит в том, что концепты или конструкты, используемые индивидуумом для

восприятия объектов или событий, влияют на его поведение. В ходе психотерапии часто необходимо создать новые конструкты или избавиться от некоторых старых, прежде чем можно будет сделать шаг вперед по пути к исцелению.

Реп-тест разрабатывался с целью помочь клиницисту выявить некоторые важные конструкты клиента в отношении других людей. Хотя существует много способов проведения теста, включая его групповую и индивидуальную версию, он всегда предполагает сортировку стимулов тем или иным способом. Реп-тест дает данные, которые можно представить в виде матрицы, или решетки (*grid*), и затем провести оценку взаимосвязей между конструктами. Один из более простых вариантов Реп-теста поможет проиллюстрировать его характерные особенности.[1] В этом варианте респонденту вначале дается перечень названий ролей (*Role Title List*) и его просят указать человека, соответствующего каждой из этих ролей в его собственной жизни. В число типичных ролей могут быть включены, например, ваш отец, ваша жена или подруга, учитель, которого вы любили, и человек, с которым вы до последнего времени близко общались, но который, как вам кажется, не любит вас. Затем проводящий тестирование выбирает из числа названных лиц троих и спрашивает: «В каком *важном отношении* двое из них похожи друг на друга, но отличаются от третьего?» Эта процедура повторяется с множеством других наборов из трех персонажей, в которых ряд персонажей может неоднократно встречаться в разных сочетаниях по три. После того как первичные данные такого рода собраны и представлены в виде решетки, можно провести анализ их импликаций. Несмотря на то что Реп-тест дает богатые качественные данные, было разработано множество количественных способов определения того, насколько важными могут быть выявленные конструкты для конкретного человека. Эти способы варьируют от простой описательной статистики до весьма сложного структурного анализа. Имеется также программное обеспечение содержательного и структурного анализа протоколов (см., например, Winter, 1992).

В своих разнообразных модификациях Реп-тест использовался в большом количестве исследований по проблемам, имеющим отношение к теории личности, социальному познанию, образованию и общению, а также к психотерапии и психологической оценке. Среди показателей, выводимых из классификаций респондентом знакомых и близких ему людей, один получил название *когнитивная сложность (cognitive complexity)*. Этот показатель основан на числе различных конструктов, используемых индивидуумом, и рассматривается в качестве меры когнитивного стиля. Более высокая степень когнитивной сложности означает, что данный человек использует больше измерений (*dimensions*) и, следовательно, пользуется более дифференцированной когнитивной системой при организации и репрезентации окружающей его среды (Bieri, 1971; Bieri et al., 1966; Goldstein, & Blackman, 1978a, p. 483–487; 1978b). Однако, когнитивная сложность сама является многоаспектным конструктом, и уровень ее корреляции с другими переменными в значительной степени зависит от того, как она концептуализируется и измеряется (см., например, Goldsmith, & Nugent, 1984).

После периода относительного затишья, продолжавшегося до 1980-х гг., интерес к теории личных конструктов Келли вспыхнул с новой силой, и с этого времени ис-

[1] Что касается других вариантов проведения теста, см. Bannister, & Mair (1968), Beail (1985), Landfield, & Epting (1987), G. J. Neimeyer (1989), Winter (1992). (На рус. яз. см.: *Франселла Ф., Баннистер Д.* Новый метод исследования личности: Руководство по репертуарным личностным методикам: Пер. с англ. — М.: Прогресс, 1987. — *Примеч. науч. ред.*)

следования в рамках этого теоретического подхода продолжали набирать темп (Bannister, 1985; Burr, & Butt, 1992; Epting, & Landfield, 1985; Fransella, & Thomas, 1988; G. J. Neimeyer, & R. A. Neimeyer, 1990; R. A. Neimeyer & G. J. Neimeyer, 1992). Теория Келли к тому же использовалась в качестве одной из опор конструктивизма.[1] Поскольку методика Келли идеально подходит для выявления тех неповторяемых способов, какими индивидуум извлекает из опыта и организует смысловые значения, в последние годы Реп-тест сам стал предметом обширных исследований. Сейчас для анализа данных репертуарных решеток исследователи применяют более сложные и тонкие статистические методы, такие как иерархический кластерный анализ и многомерное шкалирование (см., например, Merluzzi, 1991; Ogilvie, & Ashmore, 1991). Эта методика, однако, имеет столь много вариантов, что просто невозможно сделать какие-либо обобщения об ее эффективности и психометрических качествах.

Воспринимаемая среда и социальный климат. Понятие среды (*environment*) проникает в психологию с многих сторон.[2] Мы уже обсуждали важную роль оценок среды наблюдателями в связи с тестированием особых популяций (глава 9) и рассматривали воздействие культурных и ситуационных переменных на проявление когнитивных и аффективных черт (главы 9, 11, 12 и 13). С феноменологической точки зрения, оценка различных сред и социального климата путем анализа восприятий (*perceptions*), сообщенных лицами, находящимися в каждой среде, может к тому же внести существенный вклад в понимание отдельных людей и групп.

Для описания и оценки физических и социальных аспектов среды был разработан ряд измерительных инструментов (критический разбор некоторых из них см. в работе Walsh, & Betz, 1995, chap. 11). Десять Шкал социального климата (*Social Climate Scales*), разработанные Рудольфом Мусом (Moos, 1974, 1993a, 1993b, 1993c, 1994a; Moos, & Spinrad, 1984) в Стэнфордском университете, входят в число наиболее часто используемых и универсальных инструментов этого типа. Шкалы социального климата применимы к таким видам непосредственного окружения (*contexts*), как лечебные программы госпитализированных и живущих на попечении общины психически больных, приюты и исправительные учреждения, армейская обстановка, университетские студенческие общежития, классы средней школы, рабочая среда и семья. Кроме того, в число этих шкал входят более общие шкалы оценки групповой среды для проблемно-ориентированных групп, социальных групп и групп взаимной поддержки. Большинство этих шкал состоят из 90 или 100 пунктов с ответами типа «верно—неверно», с помощью которых респондент описывает свое восприятие данной среды. Формулировки пунктов первоначально составлялись таким образом, чтобы «простучать» теоретически выбранные измерения (*dimensions*) среды, такие как принудительное вовлечение (*press toward involvement*), автономность или порядок. Окончательные формулировки отбирались эмпирически, исходя из их способности диффе-

---

[1] Конструктивизм представляет собой активное и вызывающее много споров направление в психологии, которое рассматривает людей как проактивных деятелей (*agents*), стремящихся к осмыслению своего опыта, и провозглашает множественность взглядов на (по)знание — помимо прагматизма (Gergen, 1985; Mahoney, 1991; G. J. Neimeyer, 1993; R. A. Neimeyer, & Mahoney, 1995).

[2] Фактически, за три десятилетия своего развития учение об окружающей среде и поведении превратилось в междисциплинарную область исследований, быстро разрастающуюся не только в Америке, но и в других частях света (см., например, Bonnes, & Secchiaroli, 1995; Groat, 1995; McAndrew, 1993; Stokols, 1995; Stokols, & Altman, 1987).

ренцировать разные среды, а также на основе внутренней согласованности пунктов в рамках подшкал. Каждая из 10 шкал пересматривалась и обновлялась не менее одного раза (Moos, 1994b; Moos, & Moos, 1994; Trickett, & Moos, 1995). Шкалы социального климата могут предъявляться в трех разных формах, которые измеряют: а) актуальные восприятия реальной среды; б) восприятия идеальной среды; в) ожидания в отношении незнакомой среды определенного типа.

Каждое непосредственное окружение (*context*) характеризуется 7–10 показателями подшкал, которые оценивают различные измерения, или параметры данной среды. Некоторые из этих измерений повторяются в шкалах для разных видов непосредственного окружения. Фактически, несмотря на широкое разнообразие видов ближайшего окружения, охватываемых Шкалами социального климата, подшкалы для каждого вида окружения полностью укладываются в одну и ту же троичную классификацию, включающую: а) параметры отношений (например, Вовлеченность, Поддержка, Сплоченность членов своего круга); б) параметры личного роста (например, Независимость, Целеустремленность, Соперничество); в) параметры изменения и системы поддержки (например, Порядок и организация, Ясность, Новшество). Примечательно, что Шкалы социального климата разрабатывались для оценки относительно мелких единиц внутри сложных и разнородных образований, например, класса, а не всей школы; лечебной программы, а не всей больницы; университетского общежития, а не всего университета. В этом отношении они дают легче интерпретируемые и менее неопределенные данные, чем те, которые могли бы быть получены в результате комбинированной оценки целой организации. Хотя в *Руководстве пользователя (User's Guide)*, входящего в комплект Шкал социального климата, указаны такие области их применения, как клиническая оценка, организационное консультирование и оценка программ, некоторые критики считают, что эти шкалы более подходят для исследования детерминант и следствий различий в восприятии среды (Allison, 1995; Loyd, 1995; R. O. Mueller, 1995; Saudargas, 1989; Sheehan, 1995; C. R. Smith, 1989).

Отчеты наблюдателей

Обсуждавшиеся до сих пор тесты дают достаточное представление о разнообразии подходов к оценке личности. Однако по поводу большинства из них в лучшем случае можно сказать, что они являются перспективными экспериментальными методиками, пригодными для исследовательских целей, или же полезными измерительными инструментами при условии, что полученные с их помощью данные интерпретируются опытным клиницистом, увязывающим их с другой информацией об индивидууме. К настоящему времени стало очевидным: при оценке личности мы не можем всецело полагаться на стандартизованные тесты. Другие источники информации необходимы нам для того, чтобы развивать или дополнять «наводки» (*leads*) тестовых показателей, оценивать черты личности, для которых нет пока адекватных тестов, и добывать критериальные данные для разработки и валидизации личностных тестов.

Непосредственное наблюдение за поведением играет существенную роль в оценивании личности, будь то в клинике, в консультационном центре, в классе, в отделе кадров или в любой другой ситуации, требующей индивидуальных оценок. Чтобы показать место подобных наблюдений за поведением в правильной перспективе, напомним, что все тесты сами по себе являются оценками малых выборок поведения.

Конечно, эти выборки получают и оценивают в стандартизованных условиях. Но очевидные достоинства таких стандартизованных методик мы должны взвесить относительно достоинств гораздо более обширного выборочного изучения поведения благодаря доступности методик наблюдения в естественной обстановке. Если взять предельный случай, то, имей мы подробнейшую биографию какого-либо человека от рождения до 30 лет, нам, вероятно, удалось бы предсказать его последующее поведение с большей точностью, нежели при использовании любого теста или тестовой батареи. Такую запись всех мелочей и обстоятельств его жизни получить практически невозможно, но если бы она все-таки была, то мы могли бы делать прогнозы на основе 30-летней выборки поведения, а не на основе одно- или двухчасовой выборки, обеспечиваемой тестами.

Во всех рассматриваемых в этом разделе методиках информация о том, что делает конкретный человек в естественном окружении на протяжении относительно продолжительных периодов времени, передается через посредство одного или нескольких наблюдателей. Многое делается, чтобы повысить точность и полноту передачи таких наблюдений.

Натуралистическое наблюдение. Методики непосредственного наблюдения за поведением в естественной обстановке наиболее широко использовались специалистами по детской психологии, особенно в работе с детьми дошкольного возраста. Хотя подобные методики применимы к лицам любого возраста, чем младше наблюдаемый, тем менее вероятно, что на его поведение повлияет присутствие наблюдателя или что он покажет «социальный фасад» (*social facade*), затрудняющий интерпретацию поведения. Такие методики наблюдения оказались также полезными в классе, особенно если наблюдателем выступает сам учитель или кто-то еще, кто естественно вписывается в обычную школьную обстановку. Важное применение методик оценки этого типа можно найти в программах модификации поведения, осуществляемых в школах, семьях, детских садах (*child care centers*), клиниках (*clinics*), больницах (*hospitals*) или любых других условиях (Hartmann, & Wood, 1990; Kent, & Foster, 1977; Lalli, & Goh, 1993). Несколько оригинальных приложений натуралистического наблюдения было придумано для социальной психологии (Webb, Campbell, Schwartz, Sechrest, & Grove, 1981) и кросс-культурных исследований (Bochner, 1986).

Метод натуралистического наблюдения представлен широким разнообразием конкретных методик (Adler, & Adler, 1994; Jones, Reid, & Patterson, 1975; Sattler, 1988, chap. 17), начиная от комплексных, долговременных процедур, примером которых может служить дневниковый метод, и кончая узконаправленными, более короткими и лучше контролируемыми способами наблюдения, такими как методика временной выборки (*time sampling*). Временная выборка включает репрезентативное распределение коротких периодов наблюдения. В зависимости от характера и целей наблюдения такие периоды могут варьировать от долей минут до нескольких часов, причем наиболее часто используемыми являются 5-минутные или еще более короткие периоды. Сами наблюдения могут быть сконцентрированы в рамках одного дня или проводиться с интервалами в несколько месяцев. Они могут охватывать все поведение, имеющее место во время заданного периода, но чаще ограничиваются каким-то конкретным видом поведения, таким как речь, локомоция, общение или агрессия. Контрольные перечни (*checklists*) того, что следует искать, оказывает существенную помощь при ведении наблюдения. К другим вспомогательным средствам относятся графики на-

блюдений, бланки протоколов, системы кодирования и автоматические регистрирующие устройства (W. W. Tryon, 1985, chap. 7, 8). Когда это практически целесообразно, для регистрации наблюдаемого поведения можно воспользоваться магнитофоном, кино- или видеокамерой.[1] Сейчас, кроме того, имеются в наличии портативные микрокомпьютерные системы, расширяющие возможности сбора и анализа данных наблюдения (Kratochwill, Doll, & Dickson, 1991, p. 137–141; Repp, & Felce, 1990).

Можно также отметить, что натуралистические наблюдения имеют много общего с ранее обсуждавшимися ситуационными тестами, но при этом принципиально отличаются в двух отношениях: при натуралистических наблюдениях никак не контролируется стимульная ситуация и, по крайней мере, для большей части методик наблюдения характерно использование более обширной выборки исследуемого поведения. Интерес к исследованиям с помощью методик натуралистического наблюдения неуклонно растет, особенно в связи с их экологической валидностью и полезностью при оценке изменений во времени (см., например, Barkley, 1991; Kaminer, Feinstein, & Seifer, 1995).

Интервью. Следует также упомянуть об освященном веками источнике информации, пополняемом методами интервьюирования. Интервьюирование служит разнообразным целям в клинической психологии, консультировании, психологии персонала и образовании. Обсуждение методик, приложений и эффективности интервьюирования, а также исследований процесса интервью можно найти во многих литературных источниках.[2] По форме интервью могут варьировать от высокоструктурированных (не отличающихся по сути от зачитывания вслух опросника), через спланированные или направленные интервью, охватывающие заранее выбранные области, до ненаправленных и глубинных интервью, в которых интервьюер просто выступает в качестве «постановщика» (*set the stage*) и поощряет интервьюируемого говорить как можно непринужденнее. Применение структурированных интервью для клинических и исследовательских целей в области психиатрической диагностики стало уже привычным делом. Эти инструменты стандартизуются и, как правило, предусматривают получение количественных показателей в добавление к диагностическим ярлыкам; поэтому их нужно оценивать по тем же психометрическим стандартам надежности и валидности, которые применяются ко всем тестам. Критический разбор протоколов различных структурированных интервью можно найти в работах Hodges, & Zeman (1993), Kamphaus, & Frick (1996, chap. 12) и Rogers (1995).

Интервью дают, главным образом, два вида информации. Во-первых, они предоставляют возможность для непосредственного наблюдения довольно ограниченной выборки поведения, проявляемого в ситуации интервьюирования. Например, можно

---

[1] Необыкновенно полным примером оценочной системы, почти целиком основанной на данных наблюдения, служит система, разработанная Готтманом (Gottman, 1994, 1996), для анализа процессов супружеских отношений. Эта система, которая включает элементы, напоминающие методологию оценки в центрах, разрабатывалась и испытывалась в ходе долгосрочной программы исследований, нацеленной на выделение факторов, предсказывающих возможный исход брака.

[2] Обширная литература по интервьюированию включает периодические обзоры соответствующих исследований (Eder, Kacmar, & Ferris, 1989; Graves, 1993; Groth-Marnat, 1990, chap. 3; Landy et al., 1994; McDaniel, Whetzel, Schmidt, & Maurer, 1994), а также инструкции и рекомендации по совершенствованию методик интервьюирования, особенно в клинической (Bierman, 1990; Lukas, 1993; Morrison, 1995; Rogers, 1995; Shea, 1998) и кадровой (Fear, & Chiron, 1990; Webster, 1982) работе. Модель обучения базисным навыкам, применимым в большинстве ситуаций интервьюирования, можно найти в работе Gorden (1992).

отмечать речь индивидуума, его владение языком, манеру держать себя и общаться с незнакомым человеком. Второй — и, возможно, более важной — функцией интервьюирования является выявление фактов, относящихся к истории жизни человека. Прошлые действия и поступки конкретного человека — хороший показатель того, что он может сделать в будущем, особенно если они интерпретируются в свете сопутствующих обстоятельств и комментариев этим человеком своих поступков. Интервью должно касаться не только того, что происходило с данным человеком, но также восприятия им этих событий и их оценок с его стороны в настоящий момент.

От интервьюера этот метод требует высокой квалификации в получении и интерпретации данных. Интервью может привести к ошибочным заключениям, если не удалось выявить важные сведения или если полученные данные были неадекватно или некорректно интерпретированы. Критическим условием успешности интервьюера является чувствительность к идентифицирующим признакам в поведении интервьюируемого или в сообщаемых им фактах. Такие признаки подсказывают направление дальнейшего зондирования других фактов, которые или подтверждают первоначальное предположение, или опровергают его.

Рейтинги. Хотя рейтинги можно использовать во многих ситуациях и для различных целей, в настоящем разделе рассматривается использование рейтингов как оценки индивидуума лицом, высказывающим это мнение, на основе совокупных неконтролируемых наблюдений в повседневной жизни. Подобные субъективные оценки отличаются от натуралистических наблюдений тем, что данные накапливаются случайно и бессистемно; к тому же, рейтинги скорее содержат в себе интерпретацию и суждение, нежели простую регистрацию наблюдений. Однако по сравнению с натуралистическим наблюдением и интервью, рейтинги охватывают более продолжительный период наблюдения и отражают информацию, полученную в условиях, приближенных к реальным. Рейтинги широко используются при оценивании конкретных лиц в системе образования и в промышленности, при сборе критериальных данных для валидизации тестов и во многих других исследовательских целях. После 1970-х гг. существенно возросло количество исследований способов получения субъективных оценок, причем основной упор в них делается на получении всесторонних, систематических сведений и достаточной стандартизации формулировок и процедур для повышения сопоставимости данных разных исследований (Borman, 1991; Landy, & Farr, 1983; Ozer, & Reise, 1994, p. 370–371; Saal, Downey, & Lahey, 1980; Sulsky, & Balzer, 1988).

Многое можно сделать, чтобы повысить точность рейтингов. Обычно трудности возникают из-за неопределенности названий черт (свойств, качеств и т. д.), единиц шкалы или того и другого вместе. Чтобы решить эту проблему, каждую черту следует определять в точных и конкретных терминах, а субъективные оценки должны выражаться в форме, которая бы одинаково интерпретировалась всеми оценщиками. Вместо использования чисел или общих описательных характеристик, в которые различные оценщики вкладывают разный смысл, степень выраженности той или иной черты лучше определять через тщательно сформулированные поведенческие эталоны, или привязки (Dickinson, & Zellinger, 1980). Получены также данные, что относительная точность шкал разного формата может варьировать вместе с характером работы или выполняемой функцией, которые подлежат оцениванию (Borman, 1979; J. M. Feldman, 1986).

Одним из условий, влияющих на валидность рейтингов, является степень *релевантной связи (relevant contact)* оценщика с оцениваемым человеком (Freeberg, 1969; Landy,

& Farr, 1980; Paulhus, & Bruce, 1992; Wiggins, & Pincus, 1992, p. 493–496). Недостаточно просто долгое время знать человека; оценщик должен был иметь возможность наблюдать его в таких ситуациях, где могло проявиться изучаемое поведение. Например, если у работника не было возможности принимать решения в ходе выполнения своих функциональных обязанностей, эта способность не может оцениваться его непосредственным начальником. Во многих ситуациях проведения рейтингов желательно оставить время для проверки оценок другими способами, если оценщик не имел возможности наблюдать конкретное свойство у данного человека.

Рейтинги, подобно всем субъективным оценкам, подвержены ряду ошибок.[1] Хорошо известный пример — *гало-эффект.* Традиционно этот феномен определялся как склонность оценщиков поддаваться чрезмерному влиянию какой-то одной приятной или неприятной черты, окрашивающей все их суждения о других чертах индивидуума. Наличие гало-эффекта обычно выводится из наблюдаемых интеркорреляций оценок, данных разным измерениям (*dimensions*) функционирования оцениваемого человека. Хотя гало-эффект все еще рассматривается традиционным образом и считается большинством исследователей унитарным конструктом, с некоторых пор стали появляться другие возможные способы концептуализации этого феномена и, соответственно, другие подходы к его операциональному определению (Balzer, & Sulsky, 1992; Kozlowski, Kirsch, & Chao, 1986; Murphy, & Anhalt, 1992; Nathan, 1986).[2]

Другая ошибка, которая может влиять на данные рейтингов, называется *ошибкой центральной тенденции (error of central tendency)*, или тенденции использовать при оценке людей середину шкалы, избегая крайних участков. Еще одной ошибкой, которой подвержены рейтинги, является *ошибка снисходительности (leniency error)*, указывающая на нежелание многих оценщиков давать неблагоприятные оценки. В первом случае оценки группируются в центре шкалы, а во втором — в верхней ее части. Обе ошибки уменьшают рабочую область шкалы и снижают различительную способность оценок. Один из способов исключить эти ошибки состоит в использовании ранжирования или других *процедур упорядочивания оценок качества (order-of-merit procedures)*, которые вводят принудительное различение оцениваемых лиц и, следовательно, максимизируют информацию, даваемую рейтингами. Впрочем, естественно, что методики, применимые в тех случаях, когда сравнения проводятся внутри одной группы, не допускают прямых сравнений между группами, оцененными разными лицами.

Неожиданным результатом накопившихся исследований ошибок рейтинга стало признание того, что связи между мерами таких ошибок, как гало-эффект или снисходительность, и другими более прямыми показателями точности оценок, отнюдь не просты и часто противоречат интуитивным предположениям. Ряд исследователей, которые провели критический анализ литературы и метаанализ накопленных в этой области данных, пришли к заключению, что теоретические и методологические проблемы, свойственные оценке характеристик функционирования человека путем рейтингов, препятствуют использованию средств измерения таких ошибок, как меры точности рейтингов (Balzer, & Sulsky, 1992; Borman, 1991; Murphy, & Anhalt, 1992; Murphy, & Balzer, 1989).

---

[1] Большая часть исследований ошибок рейтинга публикуется в *Journal of Applied Psychology.* (Краткая, но информативная сводка типичных ошибок рейтинга дана в кн.: Психология труда: Пер. со словац. — М.: Профиздат, 1979. — С. 153–155. — *Примеч. науч. ред.*)

[2] В некоторых из этих концептуализаций гало-эффект рассматривается как зависящий, по крайней мере частично, от оцениваемого и даже специфической ситуации оценивания, а не только от оценщика.

Тем не менее качество процесса оценивания можно обычно повысить путем специального обучения оценщиков. Исследования проведения рейтингов в различных ситуациях доказали эффективность такого обучения в том, что касается повышения надежности и валидности оценок и снижения распространенных ошибок суждения (Bernardin, & Buckley, 1981; McIntyre, Smith, & Hassett, 1984; Pulakos, 1986; Stamoulis, & Hauenstein, 1993; Sulsky, & Day, 1992, 1994). Следует заметить, однако, что в программы подготовки оценщиков (судей, экспертов и т. д.) включались многие виды и способы обучения, и их результаты различаются по характеру, величине и длительности. Такая подготовка может включать снабжение оценщиков единообразными опорными эталонами оценивания, анализ распространенных ошибок рейтинга и способов минимизации их влияния, а также совершенствование навыков наблюдения. Для конкретных условий и целей проведения рейтингов оптимальным может оказаться использование какого-то одного вида подготовки или их сочетания. Однако улучшение навыков наблюдения, по-видимому, дает благоприятные результаты в большинстве ситуаций.

Клиническая оценка (*clinical assessment*) часто требует сбора данных от информантов, знакомых с формами поведения оцениваемого индивидуума. Рейтинговые шкалы обеспечивают эффективный способ сбора таких данных и особенно полезны при обследовании детей и подростков. В последние годы был опубликован ряд стандартизованных шкал для получения оценок от родителей и учителей, и теперь эти шкалы доступны для приобретения.[1] Один особенно показательный пример — Система для оценки поведения детей (*Behavior Assessment System for Children* [*BASC*] — Reynolds, & Kamphaus, 1992). Эта система включает в качестве составных частей данные самоотчета и отчета наблюдателя в дополнение к Оценочным шкалам учителя (*Teacher Rating Scales*) и Оценочным шкалам родителей (*Parent Rating Scales*), имеющимся в трех формах, перекрывающих диапазон от дошкольного до подросткового возраста (что касается рецензии, см. R. B. Kline, 1994).

Методика выдвижения кандидатур. Оценочной процедурой, особенно полезной при сборе мнений равных по положению людей, является методика выдвижения кандидатур (*nominating technique*). Впервые разработанная в социометрии (J. L. Moreno, 1953) для исследования структуры группы, эта методика может использоваться в любой группе лиц, которые находились вместе достаточно долго, чтобы познакомиться друг с другом, например в школьном классе, на небольшом предприятии, в клубе или воинском подразделении. Каждого человека просят выбрать одного или более членов группы, с которым он хотел бы учиться, работать, провести время за ленчем, играть или выполнять любую другую из перечисляемых в методике функций. Респондентов можно попросить выбрать столько членов группы, сколько они пожелают, или же назвать их в определенном порядке (первый, второй, третий выбор), или указать только одно лицо для каждой функции.

Когда эта методика используется для индивидуальной оценки, количество выборов, полученных любым отдельным человеком, может помочь распознать потенциальных лидеров (получивших много выборов), равно как и членов группы, оказавшихся в изоляции (редко или совсем не упоминавшихся). В дополнение можно рассчитать

[1] Обсуждение инструментов этого типа и критический разбор некоторых из них можно найти в работах Kamphaus, & Frick (1996), Piacentini (1993) и Witt, Heffer, & Pfeiffer (1990).

ряд индексов для более точной оценки каждого члена группы. Проще всего подсчитать сколько раз человека выбрали для выполнения конкретной функции, что можно трактовать как его внутригрупповую оценку. Методику выдвижения кандидатов можно применять по отношению к любому интересующему нас аспекту поведения. Например, респондентов можно попросить назвать человека с наиболее оригинальными идеями, или человека, на которого можно положиться в работе, или лучшего спортсмена. Кроме того, респондентов можно попросить назвать не только того, кто больше всего соответствует данной характеристике, но и того, кто менее всего соответствует ей. В последнем случае при подсчете суммарного показателя каждого члена группы положительным выборам можно было бы приписать весовой коэффициент +1, а отрицательным — весовой коэффициент –1.[1] Следует добавить, что оценки индивидуума членами его круга (*peer assessments*) можно получить и другими способами, такими как ранжирование или рейтинг, но методика выдвижения кандидатур, по-видимому, оказалась наиболее успешной и потому использовалась чаще других.

Независимо от способа, оценивание индивидуума членами его круга обычно выделялось как одна из самых надежных методик получения оценок в столь различных группах, как военнослужащие, руководители среднего звена на промышленных предприятиях, волонтеры Корпуса мира, школьники и студенты (Cole, & White, 1993; Gresham, & Little, 1993; Hughes, 1990; Kamphaus, & Frick, 1996, chap. 10; Kane, & Lawler, 1978; J. S. Wiggins, 1973/1988, p. 356–363). При проверке относительно разнообразных практических критериев межличностных отношений, такие оценки обнаружили хорошую текущую и прогностическую валидность. Эти результаты будут понятны, если принять во внимание некоторые особенности такого рода оценок. Во-первых, число людей, дающих оценку друг другу, достаточно велико и включает, в пределе, всех членов группы. Во-вторых, дающие друг другу оценку члены группы чаще всего находятся в наиболее благоприятных условиях для наблюдения за типичным поведением каждого. Они таким образом могут лучше судить о некоторых особенностях межличностных отношений, чем учителя, начальники и другие внешние наблюдатели. В-третьих, что, вероятно, наиболее важно, выборы членов группы — правильные или ошибочные — влияют на их поступки и, следовательно, хотя бы отчасти определяют характер последующих взаимоотношений каждого члены группы с остальными. Можно предположить, что и другие, сопоставимые с исследуемой, группы будут реагировать на оцениваемого индивидуума аналогичным образом. Итак, можно сказать, что социометрические оценки обладают содержательной валидностью в том же смысле, что и выборки действий (*work samples*).

Контрольные списки и Q-сортировки. Любой инструмент, основанный на самоотчетах, наподобие обсуждавшихся в главах 13 и 14 личностных опросников и инвентарей интересов, может использоваться наблюдателем при описании других лиц.[2] Средства измерения, предназначенные для оценки Я-концепции, особенно подходят для

---

[1] В тех случаях, когда допускаются отрицательные номинации, следует быть особенно осторожным, чтобы предотвратить любые потенциально вредные воздействия методики выдвижения кандидатур на участников исследования. Обсуждение этических вопросов проведения оценки детей и подростков их знакомыми сверстниками см. в Gresham, & Little (1993, p. 174–175) и Kamphaus, & Frick (1996, p. 201–203).

[2] Личностный опросник для детей (*PIC*) и Личностный опросник для юношества (*PIY*), охарактеризованные в главе 13, являются, фактически, параллельными инструментами в а) форме стандартизованного самоотчета (*PIY*) и б) в форме инвентаря наблюдаемого поведения (*PIC*).

этих целей. Контрольный список прилагательных (*ACL*) широко использовался для получения оценок наблюдателей в исследовательской программе *IPAR* (Gough, & Heilbrun, 1983). Специально обученные психологи, которые наблюдали за участником программы в течение двух- или трехдневного периода обследования, отмечали свои оценки, подчеркивая соответствующие прилагательные из контрольного списка.

Q-сортировка также широко использовалась для оценок наблюдателя. Дж. Блок (J. Block, 1961/1978) первоначально разработал Калифорнийскую колоду карт для Q-сортировки (*California Q-Sort Deck*) в целях обеспечения стандартного языка для всесторонней оценки личности профессионально подготовленными наблюдателями. Эта колода потом издавалась для более широкого распространения и неоднократно пересматривалась для того, чтобы сделать ее язык пригодным как для профессиональных пользователей, так и для пользователей-непрофессионалов. Имеется также ее адаптация — *California Child Q-Set* — для использования в работе с детьми младшего возраста (что касается рецензии, см. Heilbrun, 1985). Во всех формах стимульный материал состоит из 100 карточек с утверждениями, которые нужно разложить на 9 кучек в строгом соответствии с оговоренным количеством карточек в каждой из них. Утверждения сортируются с точки зрения их «заметности» (*salience*) у оцениваемого лица, т. е. их важности в определении его неповторимых и существенных характеристик. Таким образом сохраняется типичная для *Q*-сортировок ипсативная система отсчета, при которой индивидуум не сравнивается с внешними нормативными критериями.

Наличие таких унифицированных наборов для *Q*-сортировки облегчает сообщение и сравнимость данных, полученных разными наблюдателями. Стандартный комплект *Q*-сортировки можно также применять при решении ряда других исследовательских задач (см., например, Caspi et al., 1992; Reise, & Oliver, 1994; Wink, 1992). Еще одно применение этой методики связано с индивидуальным оцениванием. В этой связи Дж. Блок (J. Block, 1961/1978) приводит примеры трех «определяющих *Q*-сортировок», представляющих согласованное оценивание нормального индивидуума и двух психиатрических синдромов, с которыми можно сравнить *Q*-сортировку этого человека. Аналогичные определяющие *Q*-сортировки можно разработать для любой интересующей исследователя категории людей.

Биографические сведения

При обсуждении в этой главе методик интервьюирования указывалось на важность сведений из истории жизни человека. Информация о прошлом поведении и опыте представляет интерес как для теоретиков, пытающихся понять принципы развития личности и познавательной сферы человека, так и для психологов-прикладников, пытающихся оценивать конкретных людей и предсказывать поведение. Этот интерес вполне оправдан, поскольку то, как конкретный человек реагировал на определенные ситуации в прошлом, является многообещающим источником информации о том, как такой человек будет реагировать на подобные ситуации в будущем.

Данные об истории жизни человека можно получить с помощью разных методов, среди которых интервьюирование и вопросники находят самое широкое применение, особенно в клинической психологии и психологическом консультировании. Дневники и автобиографические документы также служат богатым источником сведений для психобиографов и других специалистов, заинтересованных в изучении жизни

отдельных людей (см., например, J. S. Wiggins, & Pincus, 1992, pp. 487–493).[1] А ученые, проводящие лонгитюдные исследования, не только собирают, но и создают документальные летописи фактов из истории жизни благодаря повторным наблюдениям и измерениям своих испытуемых на протяжении длительных отрезков времени (Funder, Parke, Tomlinson-Keasey, & Widaman, 1993).

Однако наиболее структурированный метод сбора и использования сведений об индивидуальной жизни представлен биографическими опросниками или шкалами, — называемыми теперь в совокупности средствами сбора и оценки биографических данных (*biodata measures*), — разрабатывавшимися для предсказания характеристик успешности человека в сфере производства и образования. Подобно личностным опросникам и инвентарям интересов, обсуждавшимся в главах 13 и 14, биографический опросник представляет собой инструмент типа стандартизованного самоотчета, ответы на пункты которого выбираются из двух или более предлагаемых вариантов, а не сочиняются самим респондентом. Хотя большинство вопросов обычно касаются относительно объективных и легко верифицируемых фактов, когда на их основе рассчитываются показатели и даются оценки или прогнозы, биографические шкалы — подобно структурированным интервью — должны отвечать тем же психометрическим стандартам надежности и валидности, как любой другой тест. Типичные вопросы относятся к уровню и характеру образования, опыту работы, особым умениям и навыкам, занятиям в свободное время и развлечениям. Часто выясняется отношение человека к своему прошлому, например, когда респондентов спрашивают о самых любимых и самых нелюбимых предметах в школе или что им нравилось (не нравилось) в их прежней работе.[2]

В историческом аспекте пункты биографических шкал отбирались и взвешивались путем привязки к внешнему критерию, как это делалось при конструировании опросников наподобие *MMPI* и инвентаря Стронга, обсуждавшихся в главах 13 и 14. Окончательный вариант опросника подвергался затем кросс-валидизации по тому же критерию на новой выборке. Когда все эти процедуры проводились должным образом, биографические опросники оказывались достаточно надежными предикторами результатов деятельности в самых разнообразных сферах. Биографические опросники разрабатывались относительно таких разнообразных критериев, как количество страховок, проданных страховыми агентами, текучесть среди банковских служащих, продуктивность ученых-исследователей, способность к художественному творчеству у старшеклассников и результативность обучения на курсах водолазов военных моряков. Такие опросники оказались валидными предикторами успешности работы в различных группах — от чернорабочих и «синих воротничков» до квалифицированных специалистов, представителей свободных профессий и управленцев высшего звена (Anastasi, 1979, p. 79–80; Owens, 1983). В то же время ограниченность многих биографических опросников, разрабатывавшихся для конкретного вида работ, да еще и эмпирическими методами, состоит в их непереносимости или нераспространимости на другие ситуации. Сфера их применения, в тенденции, ограничивается специфическими ситуациями и теми критериями, которые использовались при их разработке (Baird, 1985; Hunter, & Hunter, 1984).

---

[1] Использование автобиографических воспоминаний как проективного инструмента обсуждается в главе 15.

[2] Исчерпывающее описание области и характерных особенностей пунктов биографических опросников можно найти в работе Mael (1991).

Биографические опросники и шкалы продолжают оставаться предметом широких исследований. Были предприняты большие поисковые исследования рациональных и факторно-аналитических подходов к разработке шкал биографических данных (Hough, & Paullin, 1994; Schoenfeldt, & Mendoza, 1994). В отличие от эмпирического отбора и привязки пунктов к внешнему критерию, рациональный подход обычно начинается с идентификации основных релевантных конструктов посредством анализа содержания работы (*job analysis*) и аналитических обзоров соответствующей литературы эмпирического и теоретического характера. За этим обычно следует факторный анализ совокупности предварительных формулировок пунктов шкалы, по результатам которого отбираются окончательные формулировки пунктов получившихся факторных шкал. Относительные достоинства различных подходов к конструированию шкал биографических данных служат предметом постоянных споров, хотя достаточно очевидно, что каждому подходу присущи свои сильные и слабые стороны. Идеальный подход — это стратегия валидизации конструкта, позволяющая легитимировать как можно больше разнообразных факторов (Hough, & Paullin, 1994).

В добавление к традиционным подходам к созданию опросников опробуются новые методы генерирования, отбора и привязки к критерию пунктов шкал биографических данных в надежде сделать такие инструменты более универсальными и переносимыми (см., например, Rothstein, Schmidt, Erwin, Owens, & Sparks, 1990; Russell, Mattson, Devlin, & Atwater, 1990). Вильям Оуэнс (William A. Owens) проложил в этой области один из наиболее продуктивных путей исследования, используя методы кластеризации для идентификации подгрупп лиц, имеющих общий рисунок главных жизненных событий (Mumford, Stokes, & Owens, 1990; Mumford, & Stokes, 1992; Owens & Schoenfeldt, 1979). Такие данные можно затем применить для предсказания множественных критериальных характеристик. Более важно, однако, что изучение подгрупп, выделенных посредством этих методов, может привести к уровню понимания картины связанных с развитием изменений в жизни отдельных людей, который сочетает элементы идиографического и номотетического подходов (Hein, & Wesley, 1994).

Сегодня, после многих лет тщательно контролируемых исследований, шкалы биографических данных входят в число наиболее надежных и эффективных средств оценки и отбора в образовании, промышленности, управлении и других областях жизнедеятельности. Кроме того, история исследований средств сбора и оценки биографических сведений с необычайной ясностью высвечивает взаимозависимость фундаментальной и прикладной наук. А именно, методы и результаты исследований, проводимых для решения практических проблем отбора кадров, вносят серьезный вклад в формулирование теоретической системы для понимания развития паттернов поведения на всем протяжении жизни человека.

Тем не менее, надо признать, что внедрение средств сбора и оценки биографических данных имеет свои трудности, которые носят не только технический, но практический и политический характер. К главным проблемам относятся: а) озабоченность юридическими вопросами, такими как вторжение в личную жизнь и нарушение равных возможностей приема на работу, и б) подверженность шкал биографических данных и ретроспективных отчетов сознательной фальсификации и другим источникам ошибок (Henry, Moffitt, Caspi, Langley, & Silva, 1994; Lautenschlager, 1994; Trent, & Laurence, 1993). Большая часть опубликованных исследований по всем аспектам разработки и использования биографических инструментов была обобщена в *Biodata Handbook* (Stokes, Mumford, & Owens, 1994).

Часть 5

ОБЛАСТИ ПРИМЕНЕНИЯ ТЕСТИРОВАНИЯ



17 ОСНОВНЫЕ ОБЛАСТИ ПРИМЕНЕНИЯ ТЕСТОВ В НАШЕ ВРЕМЯ

Психологические тесты используются для решения самых разнообразных задач, и области их применения непрерывно расширяются. После более или менее подробного разбора показательных примеров разных типов тестов, обратимся теперь к рассмотрению вопросов, относящихся к их применению. В этой главе мы рассмотрим три основные области: образовательную, профессиональную и клиническую (включая консультирование), в которых тесты помогают выполнению многочисленных функций. В следующей, заключительной главе мы обсудим этические и социальные вопросы, связанные с практикой тестирования во всех трех областях.

Тестирование в образовании

Почти все типы существующих тестов используются в школах. Тесты интеллекта, тесты специальных способностей, комплексные батареи способностей и личностные тесты — все эти типы тестов можно найти в наборе инструментов консультанта по вопросам образования и школьного психолога. Учителям и администрации в системе образования часто приходится действовать в соответствии с информацией, полученной в результате проведения нескольких разных типов тестов. Однако некоторые их типы были специально разработаны для использования в сфере образования.[1] Именно эти тесты и рассматриваются в данном разделе. Они включают инструменты для предсказания и классификации (или распределения) в рамках строго определенных мест получения образования и широкий класс тестов учебных достижений.

Тесты достижений: сущность и назначение. Имея явное численное превосходство над всеми другими типами тестов, тесты *достижения (achievement)* предназначены для измерения воздействия теоретических и практических курсов обучения. Стало уже традицией противопоставлять тесты достижений тестам способностей (*aptitude*

[1] В *Стандартах тестирования* 1985 г. (AERA, APA, NCME, 1985) и в предложенном их пересмотре (см. главу 1) имеется глава, посвященная использованию тестов в образовании.

tests), относя к последним тесты общего интеллекта, комплексные батареи способностей и тесты специальных способностей. С определенной точки зрения, различия между тестами достижений и способностей есть различия в степени единообразия релевантного предшествующего опыта. А это значит, что тесты достижений измеряют влияние относительно стандартизованных последовательностей опыта, таких как начальный курс французского языка, тригонометрии или программирования. В отличие от тестов достижений выполнение тестов *способностей (aptitude)* отражает совокупное влияние разнообразного опыта повседневной жизни. Можно сказать, что тесты способностей измеряют результаты научения в относительно неконтролируемых и неизвестных условиях, тогда как тесты достижений измеряют результаты научения при частично известных и контролируемых условиях.

Другое различие между тестами способностей и достижений относится к их назначению. Тесты способностей служат для предсказания уровня последующего выполнения определенной функции или деятельности. Их используют для оценки степени целесообразности прохождения конкретным человеком того или иного специального курса обучения или для предсказания уровня его достижений в новой ситуации. Напротив, тесты достижений обычно представляют конечную оценку состояния индивидуума по завершении обучения. Главное значение в этих тестах придается тому, что конкретный человек способен делать в настоящий момент.

Однако нужно признать, что между применением тестов способностей и достижений невозможно провести жесткую границу. Некоторые тесты способностей могут строиться в расчете на весьма специфическое и единообразное предварительное обучение, а некоторые тесты достижений — охватывать относительно широкий и нестандартизованный образовательный опыт. Аналогичным образом, тест достижения можно использовать в качестве предиктора предстоящего обучения (и научения). По существу, тесты достижений служат тем же целям, что и тесты способностей. Например, тесты достижений по предметам, предваряющим собственно медицинскую подготовку, могут служить предикторами успешности выполнения программы медицинского факультета.

В стремлении освободиться от дополнительных значений, приобретенных терминами *aptitude*[1] и *achievement*,[2] их все чаще заменяют более нейтральным термином *ability*[3] в названиях средств оценки когнитивного поведения.[4] Любой когнитивный тест, независимо от его традиционного названия, обеспечивает выборочную проверку того, что индивидуум знает на момент тестирования, и измеряет уровень развития,

---

[1] Способность как готовность, склонность, предрасположенность (с оттенком изначальной данности и неизменности) к определенной деятельности. — *Примеч. науч. ред.*

[2] Достижение как превышение обычного уровня или как успех, победа. — *Примеч. науч. ред.*

[3] Способность как возможность что-то делать, например, различать цвета, считать или находить аналогии. — *Примеч. науч. ред.*

[4] Показательный пример изменений, происходящих в названиях тестов, — новые названия экзаменационных тестов Совета колледжей, официально введенные в 1994 г. За широко известной аббревиатурой *SAT* теперь скрывается *Scholastic Assessment Test* (Тест академической оценки), а не *Scholastic Aptitude Test* (Тест академических способностей). Новый *SAT* был перегруппирован и разбит на две составные части: *SAT-I*: *Reasoning Test* (Тест рассуждений), заменивший прежний Тест академических способностей, и *SAT-II*: *Subject Tests* (Предметные тесты), заменившие собой прежние Тесты достижений (*Achievement Tests*). Разумеется, изменения названий этих тестов сопровождались другими, более существенными нововведениями, которые будут рассмотрены в одном из последующих разделов данной главы.

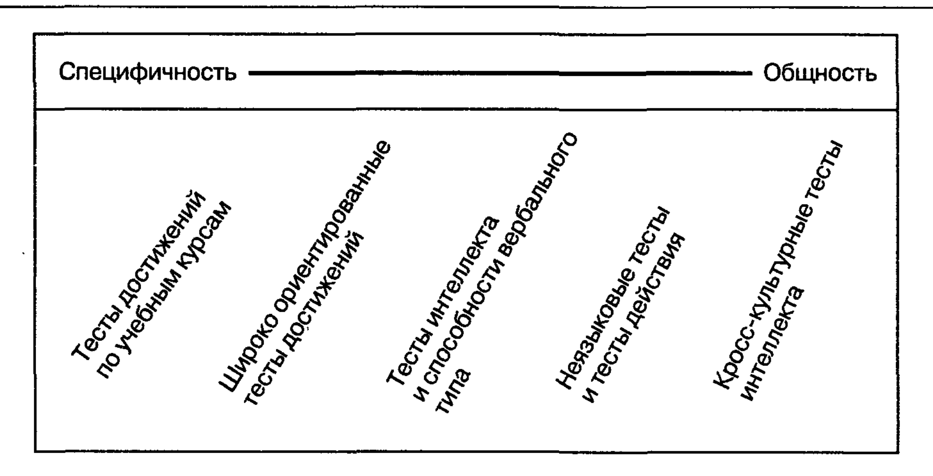


Рис. 17-1. Тесты развиваемых способностей: континуум специфичности опыта

достигнутый одной или несколькими способностями (*abilities*). Ни один тест не показывает, как или почему конкретный человек достиг такого уровня. Для ответа на эти вопросы необходимо тщательно исследовать сопутствующие переменные и особенно жизненный опыт индивидуума. В этом смысле каждый тестовый показатель имеет за собой прошлое, которое нужно досконально изучить для правильного понимания получившего его человека. Но тот же тестовый показатель имеет перед собой будущее постольку, поскольку позволяет предсказать то, как поведет себя данный человек в других, не тестовых ситуациях и к тому же по прошествии какого-то времени.

Как бы ни назывались тесты развиваемых способностей (*developed abilities*), — тестами общего интеллекта, комплексными батареями способностей, тестами специальных способностей или тестами достижений — все их можно упорядочить на континууме по специфичности жизненного опыта, предполагаемого эти тестами. Схематическое изображение этого континуума дано на рис. 17-1. На одном его конце находятся тесты достижений по учебным курсам (*course-oriented achievement tests*), охватывающие узкоспециальные умения и навыки или знание фактов. Тесты владения лексикой русского языка или навыков текущего ремонта телевизоров попали бы на этот конец континуума. Соседнее место занимают широко ориентированные тесты достижений (*broadly-oriented achievement tests*), применяемые в наше время обычно для оценки достижения главных, долгосрочных целей образования. Здесь мы обнаруживаем тесты на понимание и применение научных принципов, умение анализировать и критически оценивать художественную литературу или живопись. Еще более широко ориентированными являются тесты основных когнитивных навыков (*tests of basic cognitive skills*), — таких, как способность понимать прочитанное, умение выполнять арифметические расчеты и делать логические выводы, — которые влияют на эффективность деятельности человека в самых разных областях. Очевидно, что на этом уровне тесты достижений трудно отличить от традиционных тестов интеллекта и способностей.[1] Преимущественно вербальные когнитивные батареи, традиционно называемые теста-

[1] Это частичное перекрытие можно доказать эмпирически, на основе анализа сходства содержания тестов с такими названиями и уровня корреляции между ними (см., например, W. Coleman, & Cureton, 1954; Cooley, & Lohnes, 1976).

ми интеллекта, тесно примыкают к наиболее широко ориентированным тестам достижений. Следующими по порядку идут неязыковые тесты и тесты действия (*nonlanguage and performance tests*), обычно не требующие чтения или письма. И замыкают рассматриваемый континуум кросс-культурные тесты, предназначенные для оценки людей с самым разным происхождением и жизненным опытом.

Маркирование одних инструментов как «тесты способностей», а других как «тесты достижений» привело к ряду типичных ошибок в использовании результатов тестирования. Характерный пример — отнесение к группе «ленивых» (учащихся ниже своих возможностей) тех детей, у которых показатели по тестам достижений ниже их показателей по тестам академических способностей или тестам интеллекта. В действительности же, такие интраиндивидуальные различия в показателях тестов отражают общеизвестный факт, что никакие два теста (или другие показатели уровня выполнения, скажем отметки по учебному предмету) не коррелируют прямолинейно друг с другом. В данном случае вопрос об учении ниже своих возможностей (*underachievement*) или, наоборот, сверхдостижениях (*overachievement*) более точно может быть сформулирован как вопрос «перепрогнозирования» (*overprediction*) или «недопрогнозирования» (*underprediction*) первого теста относительно второго. Причинами ошибок предсказания в конкретных случаях являются ненадежность измерительных инструментов, различия в охвате содержания, разное влияние аттитюдных и мотивационных факторов на меры достижений и способностей, а также воздействие таких промежуточных событий, как прохождение коррективного курса обучения или длительная болезнь (R. L. Thorndike, 1963).

Уже давно различают и признают множество ролей, которые тесты достижений могут играть в образовательном процессе. Как вспомогательное средство при распределении по классам — или при любой другой оценке достигнутой компетентности — стандартизованные тесты достижений обладают преимуществами объективности, единообразия и оперативности. Если они правильно сконструированы, то обладают и другими достоинствами, такими как полнота охвата содержания и ослабление действия посторонних и случайных факторов при подсчете показателей. Тесты достижений составляют также важный элемент программ коррекционного обучения. В этой связи они могут оказаться полезными как для выявления учащихся, не способных к отдельным видам обучения, так и для измерения прогресса в ходе коррекционной работы.

Для всех типов учащихся периодическое проведение хорошо сконструированных и правильно подобранных тестов достижений может существенно облегчить процесс учения. Такие тесты выявляют недостатки прошлого обучения, задают направление последующего и мотивируют ученика. Побудительная сила «знания результатов» неоднократно демонстрировалась психологическими экспериментами во многих типичных ситуациях обучения, с различающимися по возрасту и уровню образования учащимися. Эффективность такой самопроверки обычно повышается ее оперативностью.

Рассматриваемые под другим углом зрения, тесты достижений служат средством приспособления обучения к индивидуальным потребностям. Обучение может быть наиболее эффективным лишь тогда, когда отвечает тому уровню, на котором находится ученик. Выяснение того, что ученики уже умеют делать и что они знают о предмете, есть поэтому необходимый первый шаг к эффективному обучению. Проведение тестирования в начале учебного года позволяет педагогам предпринять конструктивные шаги по ликвидации основных пробелов в знаниях учащихся, обнаруженных при выполнении тестов. Дальнейшие примеры роли тестов достижений в процессе обучения

можно отыскать в связи с предметно-ориентированным тестированием и индивидуализированными обучающими системами (см. главу 3).

Наконец, в качестве вспомогательных средств тесты достижений можно использовать для оценки и совершенствования преподавания и для формулирования образовательных целей. Тесты достижений могут дать информацию о том, какой объем знаний и навыков в действительности преподается учащимся. Привлекая внимание к таким вопросам и снабжая конкретными фактами, тесты достижений побуждают к анализу образовательных целей и содействуют критическому рассмотрению содержания и методов обучения.[1] С тех пор как повысилась подотчетность системы образования государственным органам и общественности, за несколько десятилетий было проведено беспрецедентное количество проверочных тестов в образовательных учреждениях всех уровней. В большинство случаев такое контрольное тестирование проводилось по поручению (или заказу) местных отделов образования, комитетов по образованию штата, а также федерального правительства (B. Gifford, 1989b; Hartle, & Battaglia, 1993; National Council on Educational Standards and Testing, 1992). «Государственная оценка образовательного прогресса» (*The National Assessment of Educational Progress*), называемая неофициально «табелем успеваемости нации», являет собой один из самых известных примеров непрерывной правительственной программы тестирования (см., например, Alexander, & James, 1987; Gentile, Martin-Rehrmann, & Kennedy, 1995; E. G. Johnson, 1992; Messick, Beaton, & Lord, 1983; NAEP, 1985; F. B. Womer, 1970). Поскольку применение стандартизованных тестов приобрело национальные масштабы и поскольку с их результатами могут быть связаны серьезные экономические последствия, сами эти тесты были подвергнуты усиленной проверке и критике. Методы, используемые для оценки образовательного прогресса, бывшие некогда сферой компетенции исключительно специалистов по тестированию, оказались таким образом предметом крайне политизированных дебатов, которые привлекли внимание законодателей и руководителей промышленных предприятий, а также вызвали сильный интерес у широкой общественности (R. E. Bennett, & Ward, 1993; Courts, & McInerney, 1993; Gifford, & O'Connor, 1992; S. P. Robinson, 1993; G. P. Wiggins, 1993). Обсуждение спорных вопросов и тенденций, связанных с программами тестирования и оценки школьного образования по приказам вышестоящих организаций, можно найти в работе Linn & Gronlund (1995, chap. 18).[2]

Что предпочтительнее: составление или выбор ответа? Так уж сложилось исторически, что традиционные школьные экзамены состояли из набора вопросов, на которые нужно было ответить устно или письменно. В обоих случаях экзаменуемый сам составлял и формулировал ответ. Термин *essay question* («экзаменационный воп-

---

[1] Недавние публикации, посвященные проблемам обучения математике и естественным наукам (см., например, Penner, Batsche, Knoff, & Nelson, 1993) и совершенствования мыслительных навыков (см., например, Mulcahy, Short, & Andrews, 1991), служат примером такого рода глубокой, ориентированной на принятие решений работы, являющейся следствием подобного критического рассмотрения содержания и методов обучения.

[2] На фоне грядущего реформирования школьных экзаменов в нашей стране читателям будет небесполезно познакомиться с иным взглядом на методы стандартизованной оценки учебных достижений, изложенным Джоном Равеном в его небольшой, но весьма содержательной работе (*Равен Дж.* Педагогическое тестирование: Проблемы, заблуждения, перспективы: Пер. с англ. — М.: Когито-Центр, 1999). — *Примеч. науч. ред.*

рос»)[1] стал широко использоваться для обозначения всех вопросов, предполагающих ответы в свободной форме, причем не только требующих от экзаменуемого развернутых ответов (наподобие сочинений или эссе), но и таких, на которые он должен дать короткий словесный или числовой (в виде решения математической задачи, например) ответ. В противоположность этому, «объективными вопросами» (*objective questions*) стали называть вопросы, требующие выбора ответа из предлагаемых альтернатив. Несмотря на то что есть несколько видов заданий, требующих выбора со стороны экзаменуемого, например дихотомических («верно/не верно») и на составление пар (*matching*), наиболее часто используемым, наиболее полно изученным и наиболее часто критикуемым типом тестовых заданий оказался, вне всякого сомнения, вопрос с множественным выбором ответов.

Критики формата множественного выбора утверждают, что он поощряет механическое запоминание и заучивание изолированных фактов, вместо того чтобы способствовать развитию навыков решения задач (*problem-solving*) и осмысленного понимания. В добавление к этому, многие чиновники от образования и представители политических кругов не делают различий между использованием заданий с множественным выбором и стандартизованным тестированием и огульно поносят оба этих элемента методологии оценивания.[2] По иронии судьбы, те самые программы стандартизованного тестирования, которые использовались для построения точного графика образовательного прогресса, часто обвиняются в открываемых с их же помощью недостатках образования (Courts, & McInerney, 1993; H. Gardner, 1992; Resnick, & Resnick, 1992). К сожалению, критика чрезмерного и неуместного использования стандартизованных тестов в некоторых случаях оказалась полностью оправданной. Во всяком случае, обвинения в перегруженности учебного плана проверочными тестами и заявления о необходимости срочно реформировать систему образования, включая программы тестирования, высказывались педагогами, работающими на всех уровнях этой системы, и становились все громче на протяжении последних двух десятилетий. Сторонники реформы образования считают, что нужно прежде всего пересмотреть цели учебных программ и усовершенствовать методы обучения вместе со средствами оценки его результатов, и воспринимают эти три сферы как неразрывно связанные.

Так как рассмотрение философских, политических и практических аспектов образовательной реформы выходит за рамки этой книги, мы ограничимся обсуждением ряда предложенных альтернативных методов оценки. Эти альтернативы описаны под разными заголовками: оценка «на основе анализа выполнения учебных заданий» (*performance-based assessment*), «аутентичная» (*authentic*) оценка, «прямая» (*direct*) оценка (см., например, E. L. Baker, O'Neil, & Linn, 1993; Linn, & Gronlund, 1995, chap. 10). Хотя каждый из этих подходов расставляет свои акценты в оценке учебных достижений, все их объединяет одна важная особенность — предпочтение задач, которые, подобно прежним «экзаменационным вопросам», требуют от экзаменуемого составить собственный ответ. Сейчас такие задания называют *задачами с составлением ответа (constructed-response tasks)* или *задачами со свободным ответом (open-ended tasks)*. Они противопоставляются *задачам с выбором ответа (selected-response tasks)*; это

---

[1] При кратком, терминологическом переводе на русский утрачиваются коннотативные значения, связанные со словом *essay* (эссе). — *Примеч. науч. ред.*

[2] Следует заметить, что в таких инструментах, как *SAT* и тесты *NAEP*, да и в других стандартизованных средствах измерения достижений, применяемых во многих масштабных программах образовательного тестирования, в течение некоторого времени использовались *essay questions* и другие виды заданий, предполагающих свободную форму ответа.

общий термин, применяемый к заданиям, требующим от экзаменуемого только выбрать ответ из предложенных вариантов, как это имеет место в заданиях с множественным выбором и в других типах вопросов, обычно называемых «объективными». Задания с составлением ответа могут заключаться в простом заполнении пробелов тестового бланка (*fill-in-the-blanks*), решении задач или письменном изложении вопросов и тем (*essays*), а также в разного рода демонстрациях умений и навыков, наподобие игры на музыкальном инструменте, произнесения речи или починки автомобиля.[1]

Метод, называемый *портфельной оценкой (portfolio assessment)*, предлагает другой набор альтернатив. Относящие к этому типу средства оценки нацелены преимущественно на то, чтобы сделать процесс оценивания учебных достижений как можно более обоснованным и реалистичным. Хотя этот термин применяется к широкому набору методик, портфель обычно состоит из накопленного за относительно длительный период архива выборочных образцов работ учащихся в конкретных областях, таких как письмо или любая другая деятельность, прогресс в которой можно документально зафиксировать (Camp, 1993; Gitomer, 1993; D. P. Wolf, 1993). Портфельный метод оценки предлагает пользователям гибкую стратегию и может реализовываться более или менее формально, при разной степени сотрудничества между учеником и учителем (см. работу Karlsen, 1992, которая служит примером опубликованного инструмента этого типа).

Даже из этого краткого обзора читатель, вероятно, сделал вывод, что средствам оценки работы учащихся и усвоенного ими материала уделяется огромное внимание. Интерес специалистов распространяется не только на то, что измеряют различные тестовые задания и насколько хорошо они это делают, но и на другие психологические аспекты заданий. Например, Zeidner (1993) исследовал аттитюды учащихся в отношении разных форматов тестовых заданий и обнаружил, что они предпочитают задания с множественным выбором традиционным вопросам, требующим составления ответа (*essays*). Исследование Lu, & Suen (1995) показывает, что оценка на основе анализа выполнения учебных заданий (*performance-based assessment*), в общем, ставит в более благоприятные условия поленезависимых учащихся по сравнению с полезависимыми (см. главу 16). Другие исследователи изучили взаимосвязь между тестовой тревожностью и типами заданий и пришли к выводу, что показатели по тестам с составлением ответов (*constructed-response tests*), по-видимому, больше подвержены влиянию тревоги, чем показатели тестов с выбором ответов (Crocker, & Schmitt, 1987). Обсуждение переменных, которые могут вторгаться в мотивацию тестируемых и влиять на их скорость реагирования и уровень выполнения, — таких, например, как цель процедуры оценивания, — а также некоторых других факторов, которые могут сказываться на интерпретации тестов с составлением и с множественным выбором ответов, можно найти в работе Сноу (R. E. Snow, 1993).

В то же время постепенно накапливалась эмпирическая литература, посвященная психометрическим (в строгом смысле слова) качествам задач, используемых для оценки достижений в учебных заведениях (*performance-based tasks*).[2] Темпы этих исследований, как и области получаемых результатов, широко различаются в зависимости

---

[1] Разумеется, задания теста можно классифицировать и по другим измерениям (*dimensions*), помимо измерения «составление/выбор ответа». Примеры двух разных таксономий типов заданий можно найти в работах R. E. Bennett (1993) и R. E. Snow (1993).

[2] См. в особенности R. E. Bennett, & Ward (1993).

от конкретных типов изучаемых заданий. Довольно много работ было посвящено исследованию надежности процедур определения показателя для задачи с составлением ответа, которые, подобно применяемым в ситуационных тестах процедурам (см. главу 16), часто состоят из рейтингов (E. L. Baker et al., 1993; Linn, & Gronlund, 1995, chap. 10). В общем, когда правила выставления оценок ясны и подробно разработаны, а оценщики должным образом подготовлены, получаемые коэффициенты надежности оценщика (*interrater reliability*) вполне благоприятны. С другой стороны, обобщаемость, или распространимость, результатов на другие темы и задачи оказывается типично низкой, свидетельствуя о том, что задания с составлением ответа обладают относительно высокой степенью специфичности. Этот результат не является неожиданным, учитывая, что такие задания, как правило, сложнее и допускают более широкий спектр ответов, чем задания с выбираемыми ответами.

Что касается вопросов валидности, то здесь эмпирическая база остается пока еще ограниченной, по крайней мере, в отношении наименее ограничивающих свободу экзаменуемого и наиболее новых по принципам построения типов задач с составлением ответа. Один из самых важных вопросов, требующих первоочередного решения, — это вопрос о том, в какой степени задания с составлением и с выбором ответов измеряют эквивалентные свойства, черты или навыки. Хотя данных здесь накоплено не столь уж много, обзор исследований по этой проблеме (Traub, 1993) позволяет предположить, что степень эквивалентности варьирует в зависимости от предметной области. Например, когда разные форматы заданий используются в тестах на понимание прочитанного или в тестах математических знаний, они, в общем, дают эквивалентные результаты, тогда как в области письма тип используемых заданий, по-видимому, действительно оказывает значимое влияние на тестовые показатели.

Между тем тестовые задания с множественным выбором по-прежнему широко используются в образовательных тестах. Фактически, введение различных форматов заданий в образовательные тесты, совпавшее по времени с жесткой критикой заданий с множественным выбором, по-видимому, послужило серьезным стимулом к усовершенствованию последних. Найти руководство по разработке, критическому анализу и оценке заданий с множественным выбором не составляет труда (см., например, Haladyna, 1994), а исследования конкретных аспектов этого формата, таких как оптимальное число предлагаемых вариантов ответа, продолжаются и по сей день (Trevisan, Sax, & Michael, 1991, 1994). Более того, постоянно опробуются и распространяются новые и усовершенствованные варианты задач с выбором ответа (см., например, Linn, & Gronlund, 1995, chap. 8; Sax, 1991; Sireci, Thissen, & Wainer, 1991; Wainer, & Kiely, 1987; Wainer, & Lewis, 1990).

Проводились также прямые сравнения между заданиями с составлением и с выбором ответов (например, Lukhele, Thissen, & Wainer, 1994). В большинстве случаев, при проведении сравнений по таким критериям, как экономичность, оперативность и прогностическая валидность, их результаты оказывались в пользу заданий с множественным выбором, особенно когда эти задания сравнивались с традиционными вопросами для письменного экзамена (*essay*) (Anastasi, 1988b, p. 416–418; R. E. Bennett, 1993). Нужно, однако, заметить, что проблема оценивания и сравнения различных форматов заданий для оценки учебных достижений в то время, когда и цели, и методы такой оценки находятся в состоянии непрерывного изменения, далеко не так проста, чтобы решать ее путем прямых сравнений. К тому же не следует забывать, что формат задания — это только одна из многих взаимодействующих между собой переменных, ко-

торые определяют справедливость, точность и общее качество методики оценки. Цель оценки, оцениваемая предметная область и характеристики оцениваемых лиц — все эти переменные требуют обязательного учета (E. L. Baker et al., 1993; R. E. Bennett, 1993; Dwyer, 1993; Mislevy, 1993). Например, такие вопросы, как дифференцированное влияние неудач на последующую мотивацию тестируемых могут ставиться впереди всех прочих критериев оценивания заданий, особенно для учащихся с физическими и умственными недостатками или другими особенностями, ставящими их в неблагоприятное положение. Однако стоит заметить, что в данное время нет оснований считать, будто оценка на основе выполнения учебных заданий (*performance-based assessment*) ведет к сужению разрыва, который существовал между показателями представителей белой расы и некоторых этнических меньшинств по стандартизованным тестам, построенных по принципу множественного выбора ответов. На самом деле, некоторые исследования показывают, что этот вид расхождения показателей может быть еще больше в тестах с составлением ответом, чем в тестах с выбором готовых ответов (Hartle, & Battaglia, 1993). Современные проблемы в области оценки академических достижений учащихся с выраженными культурными и языковыми различиями рассматриваются, кроме того, в работах Cancelli, & Arena (1996), K. W. Howell, & Rueda (1996), Shinn, & Baker (1996).

Типы образовательных тестов

В годовом отчете Службы тестирования в образовании за 1990 г. попечительский совет этой организации предсказал, что образовательное тестирование изменится в последующее десятилетие больше, чем оно изменилось за пять предыдущих (ETS, 1990). Похоже, что это предсказание оказалось точным и, пожалуй, могло бы быть повторено в отношении грядущего десятилетия. В настоящее время тесты всех видов подвергаются значительной переработке, и одновременно быстрыми темпами идет разработка новых оценочных инструментов. Поэтому представленный ниже обзор сосредоточен на типах инструментов, традиционно используемых в сфере образования, а не на подробной характеристике отдельных тестов. Разумеется, рассматриваются также некоторые непрерывные линии, по которым идет совершенствование этих инструментов внутри каждого типа.

Батареи общих достижений. Есть несколько батарей для измерения общих академических достижений в областях, чаще всего охватываемых учебными планами. Этот тип теста может использоваться и в первых классах, и при обследовании взрослых, хотя свое основное применение он нашел в начальной школе.[1] В типичных случаях эти батареи дают профили показателей по отдельным субтестам или в основных областях обучения. Преимущество таких батарей по сравнению с независимо разработанными тестами достижений состоит в том, что они позволяют проводить горизонтальные либо вертикальные сравнения или оба этих типа сравнений одновременно. Таким образом, относительное положение индивидуума в разных областях знаний или учебных навыков можно оценивать исходя из результатов единой нормативной выборки, а прогресс ученика от класса к классу может отображаться в единицах одной шкалы

[1] Речь идет об американской начальной школе, включающей первые 6–8 классов. — *Примеч. науч. ред.*

показателей. Пользователю теста следует выяснить, была ли выбранная им конкретная батарея стандартизована таким образом, чтобы обеспечить один из двух или оба вида сравнений.

Хотя некоторые батареи общих достижений предназначены исключительно для начальных классов, а некоторые — для средней школы, все же большинство имеют широкий диапазон, охватывающий оба уровня образования, а иногда и первый год обучения в колледже. Лишь немногие из них представлены единой для всех классов батареей, тогда как большинство состоит из нескольких частично перекрывающихся батарей, которые оформлены в виде отдельных тестовых буклетов, используемых на разных уровнях обучения. Некоторые из батарей действительно образуют согласованную серию тестов, обеспечивающую сопоставимые измерения в диапазоне от приготовительных классов (*Grades K*) до 12-го класса. Один такой набор составляют Тесты основных навыков штата Айова (*Iowa Tests of Basic Skills*), Тесты достижений и умений (*Tests of Achievement and Proficiency*) и Тесты развития в обучении штата Айова (*Iowa Tests of Educational Development*); другой — комплект Стэнфордского теста достижений (*Stanford Achievement Test Series*).

Заслуживающей внимания особенностью некоторых батарей достижений является их совместное нормирование с тестами академического интеллекта или академических способностей. Среди наиболее важных примеров — батареи достижений, сочетанные с тремя многоуровневыми тестами, разговор о которых шел в главе 10, а именно: серия Стэнфордских тестов достижений в связке с Тестом школьных способностей Отиса—Леннона; серия Тестов штата Айова и Тесты достижений и умений в связке с Тестом когнитивных способностей; Калифорнийские тесты достижений и Комплексные тесты основных навыков в связке с Тестом когнитивных навыков (см. табл. 10–1). Использование в этих случаях одной и той же выборки стандартизации дает возможность проводить прямые сравнения показателей любого ученика, полученных им по двум типам тестов. Обычно тесты каждой пары высоко коррелируют, и испытуемые получают по ним очень близкие показатели. Когда у кого-то из учеников один из показателей (либо по тестам способностей и навыков, либо по тестам достижений) значительно выше другого, желательно выяснить возможные причины такого расхождения. Батарея достижений измеряет преимущественно то, что ученик усвоил из основных школьных курсов, тогда как тест когнитивных навыков оценивает более широкий спектр умений и знаний, приобретенных учеником как в школе, так и за ее пределами. Любое значимое расхождение в выполнении этих двух типов тестов может отражать влияние специальных способностей (или, наоборот, неспособности к чему-то конкретному), либо воздействие таких некогнитивных факторов, как мотивация, интересы и аттитюды. Происхождение и жизненный опыт индивидуума часто дают подсказку к пониманию обстоятельств, вызвавших необычное расхождение в выполнении тестов.

Батареи достижений явно различаются по техническому уровню процедур, используемых для конструирования входящих в них тестов. Тем не менее как группа, эти батареи отвечают высоким стандартам разработки тестов, особенно в том, что касается объема и репрезентативности нормативных выборок, надежности и валидизации содержания. После составления заданий на основе подробной спецификации теста проводится их всесторонний анализ, включая применение методов *IRT*. Чтобы избежать смещения результатов, вызываемого половыми и этническими различиями, обычно используют специальные процедуры. Все батареи включают оценку базо-

вых навыков в таких областях, как чтение, язык и математика, и варьирующего объема предметных знаний в сфере естественных и социальных наук. Некоторые к тому же содержат ряд субтестов, предназначенных для измерения учебных навыков или умения использовать различные источники информации. Наконец, отвечая на запросы пользователей, издатели основных стандартизованных батарей достижений предлагают в настоящее время большее разнообразие заданий и опций. Теперь в них используются задания со свободным ответом и более широкий набор заданий с выбором ответа, назначение которых — измерять мыслительные навыки высшего порядка в более значимых контекстах. Издатели обеспечивают повышенную гибкость приспособления оценочных пакетов к требованиям учебных планов конкретных образовательных учреждений за счет предоставления пользователям возможности составлять разнообразные «смеси» из различных по содержанию и формату заданий и к тому же выбирать подходящую систему количественных показателей. Они также предлагают больше согласующих элементов (*linkages*) между тестами и учебными материалами.[1]

Тесты на минимум базовых навыков. Два последних десятилетия свидетельствовали о растущей озабоченности низким уровнем компетентности выпускников средней школы в таких областях, как чтение, письмо и арифметические навыки. Эта обеспокоенность привела к повышенному спросу на тесты для оценки уровня базовых навыков как средства подтверждения образовательного минимума и как основы для выдачи аттестата об окончании средней школы. Предложение сделать такое тестирование обязательным вызвало бурю споров, в которых большинство аргументов противной стороны указывало на высокую вероятность неправильного использования и истолкования тестов минимальной компетентности, а также на возможное снижение гибкости обучения и усиление бюрократических средств управления в сфере образования.[2] Хотя в большинстве штатов установлена политика в отношении тестирования минимума компетентности, определяемые ею стратегии и процедуры широко различаются в разных штатах относительно сроков проведения тестов и ступени обучения, на которой они должны проводиться; конкретного использования результатов тестирования, а также сущности и степени местной автономии в разработке или выборе тестов. Кроме того, тесты, используемые для принятия решений о выдаче или отказе в выдаче аттестата, должны отражать специфику учебного плана в разных типах школ. В силу всех этих причин такие тесты в настоящее время разрабатываются обычно самими школами, местными комитетами по образованию или соответствующими органами штата при помощи издателей тестов, которые могут предоставить технически подготовленный персонал, предложить большой банк заданий и услуги по составлению зак-

---

[1] Примером этой тенденции может служить комплект *TerraNova*, недавно изданный CTB/McGraw-Hill. В качестве составных частей в него входят: новые Комплексные тесты основных навыков (*CTBS*); вариант Множественных оценок (*Multiple Assessment edition*), который сочетает задания с выбором и задания с составлением ответа; вариант Оценок выполнения заданий (*Performance Assessment edition*), обеспечивающий более объемные задачи со свободным ответом, оценка которых может производиться либо на месте, либо издателем батареи, и, наконец, Заказной компонент (*Custom Component*), включающий дополнительные задания, предназначенные для оценки достижения целей специализированного учебного плана.

[2] Всестороннее обсуждение движения за введение тестирования минимума компетентности и порождаемых такими тестами технических психометрических проблем можно найти в Berk (1986). Еще один исчерпывающий обзор спорных вопросов и проблем, связанных с использованием тестов для удостоверения компетентности учащихся дан в Jaeger (1989).

лючений с учетом специфики местных образовательных целей. Подобные тесты могут включать некоторые компоненты обсуждавшихся выше стандартизованных батарей достижений или создаваться по индивидуальному заказу для жителей конкретного населенного пункта.

В последние годы сфера интересов, связанных с определением уровня овладения базовыми навыками, распространилась на взрослое население.[1] Совокупные последствия того, что в структуре населения значительно повысилась доля лиц с незаконченным средним образованием, а уровень компетентности выпускников средней школы снизился, вызвали озабоченность по поводу конкурентоспособности рабочей силы США на мировом рынке труда. Результаты Национального обследования грамотности взрослых (*National Adult Literacy Survey*), проведенного в 1992 г. Службой тестирования в образовании (*ETS*) под покровительством министерства просвещения США, только усилило эту озабоченность. Это обследование показало, что почти половина американцев находится на двух самых низких из пяти возможных уровней грамотности (Kirsch, Jungeblut, Jenkins, & Kolstad, 1993).

Тесты на минимум базовых навыков, специально предназначенные для взрослых, обычно разрабатываются в связи с организацией классов для обучения взрослых, реализацией образовательных программ в тюрьмах и обеспечением эффективности программ профессиональной подготовки. Замечательным примером тестов этого типа служат Тесты базового образования взрослых (*Tests of Adult Basic Education* [*TABE*, Forms 7 & 8, 1994]). Батарея *TABE* содержит пять ступенчатых уровней трудности для пяти предметных областей, включая чтение, язык и прикладную математику. Результаты тестирования сообщаются в виде статистически нормированных показателей, а также в виде основанной на квалификационных требованиях информации, которую можно использовать в диагностических целях. В добавление к стандартным формам TABE имеется в наличии специальная версия этой батареи, пригодная для использования в рабочей среде и изданная на испанском языке (*TABE Espacol*), что позволяет измерять базовые навыки испаноговорящих взрослых на их родном языке.

Тесты, создаваемые учителем для проведения в своем классе. Несомненно, что подавляющее большинство тестов на усвоение содержания конкретных учебных предметов или их разделов готовят сами учителя и используют их только в своих классах. Огромное разнообразие программ изучения одного и того же предмета, особенно в старших классах средней школы и выше, известно всем. В этих условиях никакие внешние стандартизованные тесты не могут удовлетворить потребности учителей. Однако подготовку локальных тестов для проведения в классе можно существенно улучшить, воспользовавшись методиками и опытом профессиональных разработчиков тестов. Процесс создания таких тестов можно разбить на три основных этапа: 1) проектирование теста, 2) написание заданий и 3) анализ заданий. Некоторые простые методы анализа заданий, пригодные для работы с малыми группами, описаны в

---

[1] В основе обсуждения этого вопроса лежит понятие «функциональной грамотности» (Sticht, 1975), объем которого был расширен до минимальных требований к владению языком в устной и письменной речи, пониманию и использованию различной документации и арифметическим вычислениям. Функциональная компетентность определяется исходя из требований практических ситуаций, например исходя из уровня трудности и объема материала, который должен быть прочитан для выполнения определенной работы, или, более широко, с точки зрения базовых учебных навыков, необходимых каждому человеку для налаживания собственной жизни в современном обществе.

главе 7, а краткий обзор возможных путей реализации двух других этапов дается ниже.[1]

Разработчик теста, начинающий прямо с написания заданий, скорее всего создаст односторонний тест. Без наличия технического проекта будущего теста некоторые темы изучаемого предмета могут оказаться излишне представленными в нем, в то время как другие останутся незатронутыми. Обычно по одним темам объективные задания подготовить легче, а по другим труднее. Также легче подготовить задания, которые требуют запоминания простых фактов, и труднее придумать задания на критическую оценку, обобщение различных фактов или на применение изученных принципов к новым ситуациям. Поэтому конструируемый без наличия проекта тест может оказаться перегруженным относительно недолговечным и менее важным материалом. Большинство расхожих критических замечаний в адрес тестов с выбором готовых ответов проистекает из неоправданно большого значения, которое в плохо сконструированных тестах придается механической памяти и малозначительным деталям.

Во избежание этих случайных диспропорций в охвате предметной области тестовыми заданиями, прежде чем приступать к подготовке заданий, следует составить *спецификацию теста (test specifications)* или, иначе говоря, техническое задание на разработку теста. При подготовке проверочных тестов для своего класса составление спецификации следует начинать с описания целей изучения конкретного предмета и содержания тем, подлежащих непосредственной проверке; желательно также отразить относительную важность каждого из этих аспектов в виде количества заданий, предназначающихся для проверки каждой темы и цели.[2] Разработчик теста должен также выбрать наиболее подходящую для данного материала *форму заданий (item form).* Относительные достоинства объективных и свободных заданий, обсуждавшиеся ранее в этой главе в связи с их использованием в масштабных программах стандартизованного тестирования, нужно принимать в расчет и при конструировании тестов для класса. Наконец, что касается собственно составления заданий, то на основе многолетней практики их подготовки и эмпирической оценки ответов по ним было сформулировано немало практических правил эффективного *написания заданий.* Тот, кто планирует подготовить тест для своего класса, поступит правильно, если обратится к одному из литературных источников, в которых дана сводка этих советов и указаний (например, Ebel, 1979, chaps. 4–9; Haladyna, 1994, chaps. 4–6; Linn & Gronlund, 1995, chaps. 6–9; Millman, & Green, 1989).

Тесты для университетского уровня образования. Ряд тестов и программ тестирования был разработан для использования при проведении приема, распределения по группам и консультирования студентов колледжей. Наиболее известный пример — Программа тестов академической оценки Совета колледжей (*Scholastic Assessment Tests [SAT] Program of the College Board*), состоящая теперь из двух частей, а именно, Теста рассуждений (*SAT I: Reasoning Test*), заменившего вербальный и математический разделы Теста академических способностей (*Scholastic Aptitude Test*), и Предметных тес-

[1] Более подробные указания по подготовке тестов и других методик оценки, рассчитанных на применение в рамках одного класса, см. в Linn, & Gronlund (1995, chap. 5–13).

[2] Примеры спецификации теста в табличной форме приведены в Anastasi (1988b, p. 431) и в Linn, & Gronlund (1996, p. 122). (См. также *Анастази А.* Психологическое тестирование: Пер. с англ. — М.: Педагогика, 1982. — — Кн. 2. — С. 50–51. — *Примеч. науч. ред.*

тов (*SAT II: Subject Tests*), заменивших Тесты достижений прежнего *SAT* (*SAT Achievement Tests*).[1] *SAT I* состоит, главным образом, из вопросов с множественным выбором, измеряющим вербальные и математические способности (*abilities*). Он предназначен для использования — в качестве дополнения к школьным отметкам и другой информации — при оценке готовности студента к выполнению учебной работы на уровне требований колледжа. Тесты *SAT II*, с другой стороны, предназначены для оценки знаний по конкретным предметам (например, литература, химия и всемирная история) и могут быть использованы как при распределении студентов по группам, так и при приеме.

Тесты программы *SAT* в процессе ее развития изменились по содержанию и формату заданий, а также сменили название. Например, в вербальном разделе *SAT I* усилен акцент на критическом чтении и рассуждении. Аналогично этому, некоторые задания математического раздела *SAT I* требуют теперь от студентов создавать, а не просто выбирать готовые ответы, и использование калькуляторов, согласно новым стандартам, допускается на всем протяжении работы над заданиями этого раздела теста. Тесты *SAT II* также изменились и в настоящее время включают более прямую оценку навыков по результатам опроса (*listening components*) и пробам письма (*writing samples*).

С апреля 1995 г. тестовые показатели программы *SAT* больше не выражаются в пересчете на результаты фиксированной эталонной группы, датируемые 1941 г. (см. главу 3). Вместо этого шкала показателей *SAT* была заново «центрирована» по результатам новой эталонной группы 1990-х гг., так чтобы средний уровень выполнения теста был по-прежнему представлен показателем примерно в 500 единиц.[2] Эта новая центрация сделала показатели *SAT* более точными и надежными, особенно на краях шкалы. Вдобавок ко всему, стало легче интерпретировать показатели; например, показатели по вербальному и математическому разделам теста можно теперь сравнивать друг с другом напрямую, без перевода в процентили, потому что оба они были центрированы относительно одной точки. Более того, поскольку 500 является средней точкой диапазона оценок от 300 до 800 единиц, «интуитивное» (или, иначе говоря, смысловое) и фактическое среднее будут всегда совпадать. Совет колледжей распространил специальные таблицы и другие средства, упрощающие перевод показателей оригинальной шкалы *SAT* в показатели его «перецентрированной» шкалы, чтобы сохранить преемственность между этими шкалами. Текущую информацию о надежности, уровнях трудности и процентах успешно справившихся с экзаменами по пересмотренной программе *SAT* можно найти в руководствах, специально подготавливаемых для консультантов и должностных лиц, занимающихся приемом в университеты, а также в научных отчетах и других публикациях Службы тестирования в образовании (*EST*) и Совета колледжей (College Board, 1995a, 1995b).[3] Предварительное исследование, целью которого было сравнить традиционную программу *SAT* с опытным образцом

---

[1] Что касается более подробной истории программы *SAT*, см. Anastasi (1988b, p. 328–331) и Donlon (1984).

[2] В начале 1990-х гг. средние показатели *SAT* отклонялись от 500 в обеих областях, а именно, до 424 в вербальной и до 478 в математической области.

[3] Вследствие той функции, которую экзамены по программе *SAT* выполняют при отборе студентов колледжей, сама эта программа часто становится объектом критического анализа. Недавно, например, сообщалось, что тестируемые могут правильно ответить на многие вопросы *SAT*, относящиеся к классу заданий с множественным выбором, даже не читая сопровождающих текстов к этим вопросам. Такой вывод вновь разжег угасшие было споры по поводу того, в какой степени посторонние, фоновые знания влияют на показатели *SAT* (см., например, S. Katz, & Lautenschlager, 1995).

SAT, обнаружило, что новая версия теста является несколько лучшим предиктором среднего балла (*grade point average* или, сокращенно, *GPA*) первокурсников, чем традиционная (Hale, Bridgeman, Lewis, Pollack, & Wang, 1992). Дополнительные данные о валидности экзаменов по пересмотренной программе *SAT* будут включены в техническое приложение, которое должно появиться в конце 1990-х гг.

Еще одной общенациональной программой, начатой в 1959 г., является Программа тестирования американских колледжей (*American College Testing Program* [*ACT*, 1995–1996]. Вначале ее применение ограничивалось главным образом системой университетов отдельных штатов, но программа *ACT* быстро переросла административные границы и теперь используется многими колледжами по всей стране. Современная *ACT*-оценка (*ACT Assessment*) включает четыре теста: Английский язык, Математика, Чтение (*Reading*) и Научное рассуждение (*Science Reasoning*). Отражая точку зрения ее создателя, Э. Ф. Линдквиста (E. F. Lindquist), эта экзаменационная программа предлагает тестируемому выполнить выборку основных видов работ, осуществляемых во время обучения в колледже. *ACT*-оценка частично совпадает с традиционными тестами способностей и достижений, концентрируясь на основных интеллектуальных навыках, необходимых для удовлетворительного освоения учебных программ колледжа. Некогнитивные компоненты *ACT*-оценки включают: Вопросник для сбора информации о пройденных в средней школе предметах и уровне их изучения (*High School Course/Grade Information questionnaire*), *ACT*-инвентарь интересов (*ACT Interest Inventory*) и Профиль студента (*Student Profile Section*), предназначенный для сбора сведений о стремлениях, планах и достижениях студентов, а также другой исходной информации. За всю историю своего существования *ACT* так и не достиг технических стандартов, заданных *SAT*. Однако сравнение *ACT* с другими инструментами по показателям валидности, полученным в сходных условиях, оказывается в пользу этой программы тестирования.

Следует заметить, что тесты наподобие *SAT* и *ACT* никогда не предназначались на роль заменителей школьных отметок при прогнозировании достижений во время обучения в колледже. Высокие оценки в средней школе могут предсказывать оценки в колледже не хуже или даже чуть лучше, чем большинство тестов. Однако при объединении тестовых показателей с оценками за среднюю школу, предсказание будущих успехов в освоении учебных программ колледжа улучшается. Отчасти повышение точности предсказания происходит за счет того, что единый, объективный тест выполняет функцию нейтрализации непостоянства стандартов оценивания в разных школах. Кроме того, такие тесты не подвержены воздействию возможных личных пристрастий или других случайных (связанных с произволом) факторов, которые могут влиять на выставление оценок по изучаемым в средней школе предметам.

Постепенно возрастает использование специализированных тестов достижений в качестве равносильной замены вступительных экзаменов в колледжи. Учащиеся средних школ с дополнительной подготовкой в определенных областях могут пройти тестирование по принятой Советом колледжей Программе опережающего отбора (*College Board's Advanced Placement Program* [*APP*]) для того, чтобы попробовать поступить в колледж на основании особых успехов в изучении одного или нескольких предметов. Проявление близкой тенденции можно обнаружить в Программе экзаменов университетского уровня (*College Level Examination Program* [*CLEP*]), также проводимой Советом колледжей. Главное назначение этой программы — облегчить получение

«кредитов» колледжа по набору преподаваемых в нем курсов на основе результатов тестирования[1] и обеспечить национальную систему оценивания образования университетского уровня, которое было получено путем самообразования или другими нетрадиционными методами. Аналогичная серия тестов — Квалификационные экзамены *ACT* (*ACT Proficiency Examination Program*) — проводится колледжами, признающими *ACT*-программу. Хотя эта серия тестов включает вопросы по некоторым академическим дисциплинам, таким как анатомия, физиология и патопсихология (*abnormal psychology*), она охватывает главным образом профессиональные области, такие как сестринское или бухгалтерское дело.

Прием в аспирантуру. Практика тестирования желающих поступить в колледж постепенно была распространена на аспирантуру и профессиональные школы. Большая часть созданных для этой цели тестов представляет собой комбинацию тестов общего интеллекта и тестов достижений. Широко известный пример — Письменные экзамены для аспирантов (*Graduate Record Examinations* [*GRE*]). Эта серия тестов была создана в 1936 г. в рамках совместного проекта фонда Карнеги для развития преподавания (*Carnegie Foundation for the Advancement of Teaching*) и аспирантур четырех университетов. Значительно расширившаяся с того времени, программа *GRE* проводятся теперь Службой тестирования в образовании (*ETS*) под общим руководством Совета GRE. Тестирование студентов проводится в специально предназначенных для этого центрах, созданных более чем в 100 странах мира, до зачисления в аспирантуру. Результаты теста используются университетами как вспомогательное средство при принятии решений о приеме и распределении по кафедрам, отборе кандидатов на получение стипендий и грантов, выделении научного оборудования и т. д. *GRE* включают Общий тест (*General Test*) и Предметные тесты (*Subject Tests*) в разных областях специализации.[2] В настоящее время Общий тест дает отдельные показатели вербальных, количественных и аналитических способностей. Предметные тесты имеются теперь в 16 областях специализации, включая биологию, вычислительную технику, французский язык, математику, музыку, политологию и психологию. Психометрические характеристики *GRE* приведены в самом свежем издании руководства по их проведению (*GRE 1995–96 Guide*). В общем, показатели Предметного теста являются более надежными предикторами среднего балла (*GPA*) аспирантов первого года обучения по сравнению с комбинированными показателями Общего теста или средним баллом студентов-выпускников (*undergraduate GPA*), однако комбинация всех трех мер дает наивысшие показатели прогностической валидности. Величина коэффициентов множественной корреляции, используемых в качестве таких показателей, варьирует в пределах примерно от 0,45 до 0,60 с небольшим для различных областей специализации.

В октябре 1992 г. программа *GRE* начала проводить компьютеризованную версию традиционной формы Общего теста, а в ноябре 1993 г. был введен компьютеризованный адаптивный вариант этого теста. Несмотря на первоначальные трудности, связанные с потенциальным риском получения доступа к закрытой информации при таком

---

[1] Речь идет о конкретной форме экстерната. — *Примеч. науч. ред.*

[2] До 1982 г. Общий тест назывался Тестом пригодности (*Aptitude Test*), а Предметные тесты — Углубленными тестами (*Advanced Tests*). Как и в случае с *SAT*, названия поменялись с целью избежать неправильного понимания назначения данных тестов.

проведении Общего теста *GRE*, преимущества компьютеризации оказались настолько очевидными, что Совет *GRE* может полностью отказаться от тестирования с применением бланковых форм уже к 1999 г. Кроме того, в настоящее время Общий тест *GRE* перерабатывается с целью включения в него Теста письма (*Writing Test*) и Теста математического рассуждения (*Mathematical Reasoning Test*), а также ряда вопросов с составлением ответа («Update on the New GRE», 1995).

Диагностическое и прогностическое тестирование. В отличие от рассмотренных выше батарей общих достижений и тестов достижений в конкретных областях, обсуждаемые в этом разделе тесты предназначены для анализа сильных и слабых сторон конкретного человека в освоении какой-либо предметной области и выявления возможных причин его затруднений. Большинство этих диагностических инструментов применяется при индивидуальном обследовании, и потому их чаще всего относят к разряду клинических. Тем не менее некоторые из них представляют собой отдельные части обсуждавшихся ранее основных батарей достижений и рассчитаны на групповое проведение.

Большая часть опубликованных диагностических групповых тестов оценивают навыки чтения, владения языком и выполнения математических операций и предоставляют информацию как относительно статистических групповых норм, так и относительно содержательных критериев в каждой конкретной области. Примеры этого подхода нам дают Стэнфордский диагностический математический тест (*Stanford Diagnostic Mathematics Test*) и Стэнфордский диагностический тест чтения (*Stanford Diagnostic Reading Test*), а также Калифорнийские диагностические тесты по математике (*California Diagnostic Mathematics Tests*) и Калифорнийские диагностические тесты чтения (*California Diagnostic Reading Tests*). Издатели этих двух серий выпустили также отдельные инструменты для оценки и диагностики навыков письма. Как Стэнфордская программа оценки письма (*Stanford Writing Assessment Program*), так и Система оценки письма *СТВ* (*CTB Writing Assessment System*) используют прямые выборки письма разного стиля (например, описательного или повествовательного) и предлагают несколько вариантов определения показателей.

В связи с использованием всех диагностических тестов один момент следует подчеркнуть особо. Диагностика трудностей в обучении (*learning disabilities*) и последующее проведение программы коррекционного обучения входят в обязанности соответствующим образом подготовленного специалиста. И никакая батарея диагностических тестов не может заменить его полностью в решении этих задач. Диагностика и коррекция серьезных трудностей в обучении требуют тщательного клинического обследования индивидуума, предпочтительно междисциплинарного, включая получение дополнительных сведений о состоянии органов чувств (*sensory capacities*) и моторном развитии, состоянии здоровья и болезнях, а также подробных данных об учебе, материальных условиях жизни, семейной обстановке и возможных эмоциональных проблемах.

Хотя опросы и групповые диагностические тесты могут помочь в выявлении учеников, требующих повышенного внимания, диагностика и коррекция трудностей в обучении требуют специализированных методик. Некоторые из таких методик обсуждались в главе 9, а сам этот вопрос еще будет рассмотрен в этой главе в связи с клиническим тестированием.

Некоторые типы тестов, предназначенных для использования в образовательной среде, являются по существу прогностическими инструментами. Как таковые они выполняют скорее функции тестов способностей, чем тестов достижений. В то же время они часто имеют сходство с тестами достижений по содержанию, поскольку с их помощью обычно предсказывается уровень освоения того или иного учебного предмета. Типичным образцом данного подхода служит Прогностический алгебраический тест Орлеанс–Ханна (*Orleans-Hanna Algebra Prognosis Test* — Hanna, Sonnenschein, & Lenke, 1983). В этом тесте учащимся дают серию коротких, простых «уроков» (*lessons*) по алгебре и сразу же проверяют, что они усвоили. Таким образом, тест состоит из выборок работы (*work samples*) и предсказывает ход последующего научения тестируемых по результатам выполнения ими выборочных образцов обучающих задач. Более необычный, пока еще экспериментальный образец прогностического тестирования — тесты искусственного языка, разработанные Службой управления кадрами (U. S. Office of Personnel Management) и Министерством обороны США для предсказания способности к обучению новым языкам (Diane, Brogan, & McCauley, 1991).

Другой метод оценки, несмотря на его сугубо индивидуальную ориентацию, привлекает все большее внимание, начиная с 1980-х гг. По существу, этот подход следует процедуре «тест–обучение–тест», обычно называемой динамической или направляемой (*guided*) оценкой, и связывается с коррекционным обучением. Потенциал научения индивидуума оценивается посредством наблюдения за тем, насколько хорошо он обучается индивидуально под руководством специалиста, который выполняет функции тройной роли: экзаменатора, учителя и клинициста. Главным представителем этого подхода является Фейерстейн (Feuerstein, 1979); несколько родственных подходов рассматриваются в работах A. L. Brown, Campione, Webber, & McGilly (1992) и Lidz (1987, 1997). Вследствие ее выраженного клинического характера, динамическая оценка рассматривается более полно в соответствующем разделе данной главы.

Проведение замеров в соответствии с учебным планом (*curriculum-based measurement*) являет собой пример еще одного подхода, в рамках которого была разработана группа методик, связывающих оценивание с программами вмешательства (Deno, 1992; Fuchs, 1993; Fuchs, & Deno, 1991; Shinn, 1989). Хотя одни подходы к оценке на основе учебного плана могут быть совершенно неформальными, другие предполагают использование стандартизованных средств измерения выполнения программы учеником в том, что касается таких базовых навыков, как чтение, правописание и арифметические вычисления (сравнение разных моделей см. в работе Shinn, Rosenfield, & Knutson, 1989). Общим знаменателем всех этих методик является их выраженная поведенческая ориентация и прямая связь с задачами, составляющими типичный учебный план начального образования, в отличие от предполагающих логический вывод и ориентированных на статистические нормы традиционных психометрических инструментов. Оценка на основе учебных планов использовалась первоначально в условиях специального обучения.

Оценка обучения в раннем детстве. За последние три десятилетия было опубликовано множество новых методик для измерения результатов дошкольного воспитания и подготовки маленьких детей к обучению в школе. Ряд факторов повлиял на масштабы и характер этой деятельности (см. главы 9 и 12). Исследование когнитивного развития в раннем возрасте, быстрое распространение программ для обучения до-

школьников и широкая озабоченность влиянием культурных барьеров на способность ребенка извлекать пользу из школьного обучения[1] — все это сыграло важную роль в разработке соответствующего инструментария. Некоторые из этих тестов предназначались главным образом для измерения результатов обучения в раннем детстве и потому выполняют функцию тестов достижений. Другие представляются как прогностические инструменты для оценки готовности ребенка к обучению в 1-м классе школы. Однако эти два типа инструментов незаметно переходят друг в друга, и каждый из них обычно может выполнять обе задачи.

Школьная готовность (school readiness) указывает, по существу, на приобретение необходимых умений и навыков, знаний, аттитюдов, мотиваций и других целесообразных черт поведения, которые дают возможность ученику извлекать максимальную пользу из обучения в школе. Эти предпосылки составляют то, что Хант и Кирк (J. McV. Hunt & Kirk, 1974) называли «навыками вхождения» (*entry skills*), необходимыми ребенку для того, чтобы справиться с ситуацией обучения-учения, когда он поступает в 1-й класс. Готовность связана с минимальными уровнями физического и сенсомоторного развития, достигаемыми вследствие созревания и дошкольного воспитания. Все большее значение придается иерархическому развитию знаний и умений, в силу чего овладение простыми понятиями подготавливает ребенка к усвоению более сложных понятий в последующем.

Тесты готовности обычно проводятся при поступлении в школу. Хотя они похожи на тесты интеллекта для начальных классов, в них больше внимания уделено способностям, считающимся необходимыми для того, чтобы научиться читать. Кроме того, определенное место отводится выявлению предпосылок числового мышления и сенсомоторного контроля, необходимого при обучении письму. Среди конкретных функций часто проверяются зрительное и слуховое различение, моторный контроль, понимание на слух, словарный запас, количественные понятия и общую осведомленность. Широко используемой батареей готовности являются Национальные тесты готовности, изданные теперь уже в шестой редакции (*Metropolitan Readiness Tests, Sixth Edition* [*MRT6*] — критический разбор более ранней редакции см. в Mabry, 1995 и Stoner, 1995). Иной подход к школьной готовности иллюстрируется тестами, сконцентрированными на понимании ребенком широко используемых относительных понятий, такими как Тест базисных понятий Боэма, пересмотренный (*Boehm Test of Basic Concepts — Revised*) и Шкала базисных понятий Брейкена (Bracken Basic Concept Scale).[2] Обсуждение понятия школьной готовности вместе со спорными вопросами, касающимися ее измерения, можно найти в работе Gredler (1992).

Заключительные замечания. Сфера образования находится в состоянии непрерывного изменения, и образовательное тестирование отражает это состояние как в США, так и во всем мире. По всей видимости, происходящие здесь перемены в обо-

[1] Озабоченность влиянием культурных различий на эффективность обучения маленьких детей нашла свое высшее выражение в виде общенациональной образовательной цели, гласящей, что к 2000 г. все дети США будут поступать в школу готовыми к обучению (National Council on Education Standards and Testing, 1992).

[2] Что касается критических разборов этих инструментов, см. Fitzmaurice, & Witt (1989); Linn (1989); Turco (1989); Ysseldyke (1989).

зримом будущем не только продолжатся, но и усилятся.[1] Многие эксперты сходятся в том, что необходимо добиваться большей интеграции оценки и обучения, причем таким образом, чтобы эти стороны образовательного процесса лучше дополняли друг друга, принося пользу каждому ученику[2] (H. Gardner, 1992; Nitko, 1989). Большинство из них также признают, что и тестирование, и обучение будут непрерывно совершенствоваться по мере развития теории и эмпирических исследований. Никогда не наступит такое время, когда мы удовлетворимся каким-либо одним способом оценки или методом обучения, поскольку каждому из них присущи свои ограничения. Кроме того, разные ученики требуют разных подходов. Поэтому поиск усовершенствований в сфере образования с необходимостью будет продолжаться.

Тестирование в сфере профессиональной деятельности

Психологические тесты обычно используются в качестве вспомогательных средств при принятии решений, связанных с профессиональной деятельностью, включая индивидуальное профконсультирование и решения руководителей организаций, касающиеся подбора и расстановки кадров. В этом разделе мы обсудим приложения тестирования, которые относятся к оценке профессиональных качеств индивидуума в том виде, как эта функция выполняется организациями, ответственными за отбор, подбор, расстановку и оценивание персонала.

Соответствующие структуры торгово-промышленного сектора, правительственных учреждений всех уровней, от федерального до местного, и различных родов войск используют практически все доступные виды тестов для принятия кадровых решений. Нередко комплексные батареи способностей и тесты специальных способностей разрабатывались именно для подобного применения, как и ситуационные тесты, о которых шла речь в главе 16. Все большее применение в этой сфере находят личностные (глава 13) и биографические (глава 16) опросники. Краткий обзор использования тестов и других инструментов при отборе и распределении персонала дан в работе Landy et al. (1994), а всестороннее рассмотрение этой темы можно найти в трех современных книгах (Rumsey, Walker, & Harris, 1994; Schmitt, Borman, et al., 1993; Schuler, Farr, & Smith, 1993). Основные аспекты использования тестов в промышленно-организацинной среде глубоко проанализированы в нескольких главах многотомного справочного руководства под ред. Даннетта и Хока (Dunnette, & Hough, 1990–

[1] См., например, сборник работ под ред. Бейкера и О'Нила (E. L. Baker, & O'Neil, 1994), где даны оценки технологических инноваций в теоретическом и практическом обучении; статью Сноу и Ломана (R. E. Snow, & Lohman, 1989), в которой обсуждаются последствия развития когнитивной психологии для измерений в сфере образования, а также работу Окленда и Хемблтона (Oakland & Hambleton, 1995) о современных международных тенденциях и разработках в области академической оценки.

[2] Фактически, сейчас быстро появляются новые автоматизированные системы, сочетающие в себе обучающие, оценочные и управляющие функции и допускающие индивидуальную настройку. Два современных образца систем такого класса — программа *LearningPlus*, разработанная Службой тестирования в образовании (*ETS*) для взрослых учащихся, нуждающихся в улучшении своих базовых учебных навыков, и разработанная *IBM* компьютерная программа *SchoolVista* для всех уровней довузовского образования — от приготовительных классов до 12-го класса средней школы (К–12).

1992).[1] Общество промышленной и организационной психологии (SIOP, 1987) подготовило и приняло систему принципов валидизации и использования методов подбора кадров. Характеризуя оптимальный способ действий при выборе, разработке и оценке всех процедур отбора персонала, эти принципы весьма полезны и для стандартизованных тестов. Плюс ко всему, ныне действующие *Стандарты тестирования* (AERA, APA, NCME, 1985), как и их предложенный недавно пересмотр (см. главу 1), включают главу, посвященную тестированию при приеме на работу. Еще одной важной областью применения профессионального тестирования, которая также охвачена в обоих вариантах *Стандартов тестирования*, является выдача лицензий и свидетельств лицам, признанным подготовленными к тому, чтобы заниматься каким-либо ремеслом или какой-либо деятельностью.[2]

Как и при рассмотрении образовательных тестов, в этом разделе мы сосредоточимся на тестах, специально разработанных для целей профориентации, отбора и расстановки кадров в добавление к тем инструментам более широкого назначения, которые обсуждались в других главах. Мы также кратко рассмотрим некоторые из процедур и спорных вопросов, связанных с комплектованием и валидизацией тестов, используемых при найме рабочих и служащих.

Валидизация тестов для отбора наемных работников

Как нанимателю, так и наемному работнику представляется крайне важным, чтобы люди получали должности, соответствующие их квалификации. Эффективное распределение персонала означает также, что качества, несущественные для выполнения данной работы, не должны благоприятно или неблагоприятно сказываться на решении об отборе. Если тест механических способностей требует гораздо более высокого уровня понимания читаемого, чем этого требует определенная должность, его применение вряд ли приведет к наиболее эффективному использованию кадров для этой должности. Давно известна простая психометрическая истина, что валидность теста должна устанавливаться для конкретных областей его применения. В настоящее время она приобрела особую актуальность из-за широкой озабоченности трудоустройством представителей меньшинств, оказавшихся в невыгодном положении вследствие культурных барьеров или пробелов в образовании (см. главу 18). Невалидный или содержащий не связанные с работой элементы тест может стать причиной несправедливого отказа в приеме на работу тем представителям меньшинств, которые могли бы удовлетворительно с ней справиться.

Еще одна серьезная озабоченность, которую испытывают организации и общество в целом, вызвана подтвержденной связью между продуктивностью работы и валидностью инструментов отбора. Методы оценки этой связи и типичные результаты приводились в главе 6. Оцениваемые выигрыши и потери в производительности, связанные

---

[1] Самый последний том этого руководства, вышедший под ред. Трайандиса, Даннетта и Хока (Triandis, Dunnette, & Hough, 1994), посвящен проблемам промышленной/организационной психологии в различных культурах мира.

[2] Тестирование в сфере профессиональной деятельности, включая тестирование для выдачи свидетельств и лицензий психологам, рассматривается Анастази (Anastasi, 1988b, p. 468–474). Что касается совсем недавнего рассмотрения вопросов, касающихся обоснования и применения лицензирования и аттестации вообще, см. специальный выпуск *Evaluation & the Health Professions* под ред. LaDuca (1994).

с повышением или снижением валидности процедур отбора персонала, оказываются весьма существенными. В организациях наподобие правительственных, которые нанимают много служащих, суммарный выигрыш или совокупные потери в продуктивности труда настолько велики, что заслуживают пристального внимания.

На протяжении нескольких десятилетий в психологии персонала преобладало мнение, что применяемые для отбора тесты должны подвергаться полномасштабной валидизации относительно локальных критериев выполнения работы. Конкретные методы такой валидизации путем предсказания критерия рассматривались в главах 5 и 6. Однако в подавляющем большинстве ситуаций провести полномасштабное лонгитюдное исследование валидности таких тестов просто нереально. Даже в самых благоприятных условиях, при наличии доступа к большим выборкам наемных работников, обнаруживается несколько практических ограничений (см., например, Anastasi, 1972; J. T. Campbell, Crooks, Mahoney, & Rock, 1973). Ввиду этих практических трудностей проведения полномасштабной валидизации путем предсказания локального критерия, был опробован ряд альтернативных методов.

Общие процедуры для оценки выполнения работы. Один подход к отбору персонала использует процедуры оценки, которые максимально приближены к выполнению работы в полном объеме. Тем не менее сходство между процедурой отбора и реальной деятельностью никогда не может быть абсолютным. *Принятие на работу с испытательным сроком (probationary appointment)* ближе всего к тому, чтобы быть точной копией работы в конкретной должности. Но даже в этом случае краткость испытательного срока и знание о том, что назначение на должность является проверкой, могут сказываться на поведении работника в целом ряде отношений. *Выборочный анализ работы (job samples)* представляет собой другую попытку аппроксимации реального выполнения работы. В этом случае предлагаемая кандидату задача фактически является частью работы, выполняемой на реальном рабочем месте, однако и сама задача, и рабочая обстановка остаются неизменными для всех кандидатов. Некоторые тесты этого типа изготавливаются по специальному заказу, с учетом специфики конкретного вида работ. При этом в первую очередь исходят из соображений репрезентативности выборки образцов работы и степени сходства тестовой задачи с реальной рабочей обстановкой. Известные примеры — тесты вождения (*driving tests*) и стандартизованные тесты для оценки навыков канцелярской (конторской) работы, скажем, печатания на машинке или обращения со счетной техникой.

В некоторых тестах для воспроизведения выполняемых на рабочем месте функций используется *моделирование (simulation)*. Между моделированием и выборочным анализом работы нет четкой границы, они незаметно переходят друг в друга. Примеры варьируют от выполнения операций на миниатюрном дыропробивном прессе до управления движением электропоезда или пилотирования самолета на соответствующих имитаторах. Тренажеры-имитаторы использовались для целей тестирования и подготовки специалистов в программах *NASA* и ряда военных ведомств.

К этому перечню можно добавить *методики оценки в центрах* (см. главу 16), которые получили широкое применение при оценивании управленческого или административного персонала (Bray, 1982; Finkle, 1983; Moses, 1985; Thornton, & Byham, 1982). Отличительной особенностью этого подхода является включение ситуационных тестов, наподобие *«лотка для входящих документов» (in-basket)* — методики, применяемой для тестирования администраторов в самых разных областях деятельности

(N. Frederiksen, 1962, 1966; Shapira & Dunbar, 1980). Имитируя знакомый всем «лоток для входящих документов», закрепленный на столе администратора, этот тест предлагает испытуемому тщательно подготовленный набор поступающих писем, служебных записок, докладных, бумаг на подпись и других аналогичных материалов. Перед прохождением теста испытуемому предоставляется возможность ознакомиться с вводной информацией для того, чтобы он мог составить представление о характере гипотетической работы и сориентироваться в обстановке. Задача собственно теста состоит в том, чтобы обработать все скопившиеся в лотке материалы и решить все поставленные в них вопросы, как это пришлось бы сделать испытуемому на реальном рабочем месте. Все его действия должны фиксироваться письменно, но могут включать и деловые письма, служебные записки, резолюции, планы, приказы, получаемую или передаваемую информацию, повестки дня предполагаемых совещаний или любые другие записи. В других методиках оценки в центрах могут применяться разыгрывание ролей, групповое решение проблем и деловые игры. Их общая особенность — использование группы экспертов-оценщиков и рейтингов кандидата, полученных от лиц одного с ним положения. Большинство оцениваемых таким способом свойств и черт имеют отношение к мотивации, навыкам общения и другим личностным переменным.

Несмотря на то что эти общие процедуры для оценки выполнения работы опираются, по крайней мере отчасти, на сходство с реальной деятельностью как доказательство их «принадлежности к одному роду», сами они также подвергались, по одиночке или в сочетании, оценке относительно разнообразных критериев (см. Landy et al., 1994; Schmidt, Ones, & Hunter, 1992).

Анализ содержания работы и метод рабочих элементов. Отмечается растущий интерес к применению содержательной валидизации тестов для отбора персонала. Во всех своих формах такая валидизация опирается на полный и систематический анализ содержания работы (McCormick, 1979). Чтобы быть эффективным, анализ содержания работы (*job analysis*) должен установить требования, которые отличают определенный вид работы от всех других. Описание на языке неопределенных утверждений общего характера, одинаково применимых к большинству работ, для этой цели оказывается бесполезным. Чтобы получить достаточно полную картину конкретной профессиональной деятельности, аналитик может черпать сведения из нескольких источников информации. Он может воспользоваться опубликованными руководствами по обучению конкретной профессии или должностными инструкциями, официальными отчетами о выполнении определенных видов работ и, что особенно важно, может получить консультацию экспертов в данной области — инструкторов производственного обучения, опытных работников и их непосредственных руководителей.

Эффективный анализ содержания работы должен к тому же сосредоточиваться на тех аспектах профессиональной деятельности, которые позволяют четко различать хороших и плохих работников. В ставшей классической книге К. Халла «Тестирование способностей» (C. L. Hull, 1928) подчеркивается важность этих дифференцирующих аспектов выполнения работы. Позднее на важность этого принципа указал Дж. Фланаган (J. C. Flanagan, 1949, 1954), предложивший метод критических случаев (*critical incident technique*). По существу, данный метод требует фактографического описания конкретных образцов выполнения работы, типичных как для хороших, так и для плохих работников.

Сосредоточение на критических требованиях конкретных видов профессиональной деятельности привело к разработке метода рабочих элементов (*job element method*) для конструирования тестов и доказательства их содержательной валидности (McCormick, 1979, 1983; McCormick, Jeanneret, & Mecham, 1972; Primoff, 1975; Primoff & Eyde, 1988). Этот метод был полностью разработан и широко применялся Примовым и его сотрудниками в Службе управления кадрами США (когда-то — Комиссии по государственной гражданской службе США (U. S. Civil Service Commission)). В сущности, рабочие элементы — это единицы описания критических требований, предъявляемых конкретным видом работы к работнику. Хотя разные адаптации метода рабочих элементов различаются деталями процедуры, все они дают описание профессиональной деятельности на языке специфических требований к поведению работника, исходя из которых можно прямо формулировать задания теста. Конкретные поведенческие формулировки могут, в свою очередь, объединяться в более широкие категории, или конструкты, — такие как точность вычислений, развитая тонкая моторика, зрительное различение или способность работать под давлением (*to work under pressure*). Наблюдается рост числа исследований, нацеленных на разработку общей таксономии профессиональной деятельности на основе широких поведенческих конструктов (Fleishman, 1975; Fleishman, & Quaintance, 1984; Fleishman, & Reilly, 1992b).

Методы анализа содержания работы могут способствовать более эффективному использованию тестов отбора для множества внешне непохожих профессий. Это можно проиллюстрировать на примере таких инструментов, как Обзор Флейшмана для анализа содержания работы (*Fleishman Job Analysis Survey* [*F-JAS*]) и система «Ключевые элементы труда» (*Work Keys system*). Обзор Флейшмана представляет собой средство анализа работы в целях описания различных видов работ с точки зрения знаний, навыков и способностей, необходимых для их выполнения. Пятьдесят две из его 72 шкал охватывают тщательно определяемые способности в когнитивной, психомоторной, физической и сенсорноперцептивной областях, причем большинство этих шкал были увязаны с существующими тестами (Fleishman, & Mumford, 1991; Fleishman, & Reilly, 1992a, 1992b). Остающиеся 20 шкал имеют отношение к двум областям: межличностной/социальной (Interpersonal/Social) и знаний / навыков (Knowledge/ Skills), — и все еще находятся в состоянии развития. С другой стороны, система «Ключевые элементы труда» — совсем недавно разработанная в рамках программы *ACT* (ACT, 1995; Scruggs, 1994), — сконцентрирована на гораздо меньшем наборе общих рабочих навыков, таких как «Поиск информации», которым можно обучить в разумно короткие сроки. Однако внутри этой строго очерченной области, система *Work Keys* предоставляет сопряженный пакет средств для: 1) анализа особенностей работы (*job analysis*) и построения профиля; 2) оценки уровня навыков; 3) обеспечения обратной связи с тестируемыми, преподавателями и работодателями; 4) информационной поддержки в реализации плана практической или образовательной подготовки.

Анализ содержания работы — один из старейших и наиболее жизнеспособных методов, разработанных в промышленной психологии. Его применение при валидизации тестов для отбора наемных работников стало расширяться по мере того как прогресс в области компьютерных технологий сделал сбор и анализ данных о выполняемой работе более дешевым.[1] В добавление к этому, собранная в ходе тщательного

[1] Обзор достижений и проблем в области методологии анализа содержания работы (*job analysis*) можно найти в работе Harvey (1991). Нэпп, Рассел и Кэмпбелл (Knapp, Russell, & Campbell, 1993) описывают специфические приложения анализа содержания работы в условиях отбора и распределения персонала вооруженных сил США.

анализа работы информация может найти применение для решения множества других задач, таких как определение рыночной стоимости работы или проектирование рабочих мест (см., например, Campion, 1994; I. L. Goldstein, Zedeck, & Schneider, 1993).

Предсказание характеристик выполнения работы. Практические трудности, связанные с проведением валидизации посредством предсказания локального критерия (см. главы 5 и 6), привели к относительному дефициту таких исследований. Тем не менее фактическое положение дел таково, что многие организации нуждаются в прогнозах интенсивности и качества будущей работы для того, чтобы принимать решения по поводу отбора и расстановки кадров.[1] В качестве альтернатив для этих целей все чаще выбираются методики синтетической валидизации и обобщения валидности. Обе они позволяют оценивать валидность теста для конкретной работы, не прибегая к локальной валидизации. Кроме того, по мере накопления эмпирических данных обе методики должны давать сходящиеся результаты в отношении существа оцениваемых конструктов (J. P. Campbell, 1990a).

Понятие *синтетической валидизации (synthetic validation)* основано на предпосылке метода рабочих элементов, состоящей в том, что всегда имеется возможность выявить навыки, знания и другие требования к рабочим характеристикам индивидуума, которые являются общими для многих видов выполняемой работы. Синтетическая валидность (*synthetic validity*) была определена как «предварительный вывод о валидности в конкретной ситуации на основе систематического анализа элементов работы, определение валидности теста для этих элементов и объединение валидностей элементов в единое целое» (Balma, 1959, p. 395). По существу, эта методика включает три этапа: 1) всесторонний анализ содержания работы для выявления ее элементов и определения их относительного веса в реальной деятельности на данном рабочем месте или в данной должности; 2) анализ и эмпирическое исследование каждого теста для установления того, в какой степени он измеряет уровень выполнения каждого из выделенных на предыдущем этапе элементов работы; 3) нахождение валидности каждого теста для данной работы путем синтеза весов этих элементов в реальной работе и в соответствующем тесте. Статистическая процедура для вычисления синтетической валидности была разработана Примовым (Primoff, 1959; Primoff & Eyde, 1988). Получившая название *J*-коэффициента (от *job coefficient*), эта процедура, по существу, является адаптацией уравнений множественной регрессии, описанных в главе 6. Другие подходы к синтетической валидизации описаны в работах J. P. Hollenbeck, & Whitemer (1988) и Mossholder, & Arvey (1984).

Процедуры обобщения валидности (*VG*), — впервые разработанные Шмидтом и Хантером (Schmidt, & Hunter, 1977) и рассмотренные в главе 5, — предоставляют другую возможность для валидизации тестов отбора персонала. В своей основе этот подход позволяет применять накопленные ранее данные о валидности к новым ситуациям благодаря использованию методик метаанализа (Schmidt, Hunter, Pearlman, & Hirsh, 1985). Действительная степень обобщения данных, получаемых в результате метаанализа, была подвергнута сомнению со стороны некоторых специалистов. Критики обратили внимание на различия, существующие между ситуациями, в которых выполняется работа, и на методологические проблемы в оценивании параметров. В свою

[1] Продуктивное обсуждение ограничений предсказания как парадигмы отбора персонала см. в de Wolff (1993).

очередь, это привело к усовершенствованию метааналитических методов и, как следствие, к их возросшему признанию и применению. Несмотря на остающиеся спорные моменты и на сохраняющиеся возможности дальнейшего совершенствования, методы обобщения валидности во многом способствовали поддержанию жизнеспособности теории, исследований и практики профессионального тестирования (см., например, L. R. James et al., 1992; Landy et al., 1994; Schmidt et al., 1993).

Критерий выполнения работы. Некоторые из наиболее перспективных работ в области отбора и распределения персонала вызваны возобновившимся интересом к тому, что задается в качестве критериев. Напомним, отсылая к главе 5, что существует широкое множество индексов, которые можно рассматривать как меры критерия, в зависимости от того, как определяется критерий. Тем не менее в рамках каждого валидационного исследования обычно использовалась какая-то одна удобная мера выполнения работы, представляющая «единственный» критерий, независимо от цели процесса предсказания. До последнего времени, несмотря на неоднократные призывы к более тщательному рассмотрению критериев, раздававшиеся в течение нескольких десятилетий (см., например, L. R. James, 1973; Tenopyr, 1986; Wallace, 1965), практически не предпринималось никаких попыток решения этой важной проблемы. И только в последние годы некоторые исследователи попытались добиться более четкой концептуализации профессиональной деятельности (*job performance*) и лучшего понимания ее детерминант (Borman, 1991; Campbell, McCloy, Oppler, & Sager, 1993; B. F. Green, & Wigdor, 1991; Schmidt, & Hunter, 1992).

Одна новая модель профессиональной деятельности (*job performance*), которая обещает сыграть важную эвристическую роль в этой области, — мультифакторная теория, разрабатываемая Джоном Кэмпбеллом и его сотрудниками в связи с Проектом отбора и распределения специалистов сухопутных войск США или, как его сокращенно называют, Проектом А (J. P. Campbell, 1990a, 1990b, 1994; Campbell, McHenry, & Wise, 1990). Модель Кэмпбелла учитывает многоаспектный характер выполнения работы (*job performance*) и выделяет ее разнообразные элементы, распределяя их по соответствующим категориям. На начальной стадии эта модель проводит принципиальные разграничения между теми аспектами оценивания работы, которые контролируются работником, — например, поведение, связанное с самим выполнением работы, и теми, которые он не может контролировать, — например, последствия выполнения работы (эффективность), связанные с ней относительные затраты (производительность) и значение, придаваемое каждому из этих аспектов организацией (полезность). Что касается собственно выполнения работы, данная теория постулирует, что любая работа предполагает множество исполнительных компонентов (задач) и что детерминанты каждого компонента состоят из различных сочетаний знаний, навыков и мотиваций работника. Кроме того, каждая детерминанта выполнения работы имеет несколько предпосылок (*antecedents*), допускающих более или менее точное определение — таких, как профессиональная подготовка, непредвиденные подкрепления (*reinforcement contingencies*) и индивидуальные особенности, — которые могут косвенно влиять на характеристики выполнения через воздействие на уровень знаний, навыков и мотивации конкретного человека. Вдобавок ко всему, эти детерминанты выполнения работы взаимодействуют между собой, что также сказывается на выполнении работы.

Хотя многофакторная теория профессиональной деятельности (*job performance*) находится еще в процессе развития (J. P. Campbell, 1990a, 1994; D. J. Knapp, & Camp-

bell, 1993), ее общий замысел созвучен другим важным концептуальным и методологическим успехам в оценивании трудовой деятельности (Borman, 1991). В настоящее время данная модель идентифицирует восемь общих факторов выполнения работы, включая такие характеристики, как согласованность усилий, личная дисциплина, лидерство, опытность в отношении выполнения конкретных рабочих заданий и другие виды профессионального опыта. Предполагается, что эти факторы достаточно широки для того, чтобы охватить главные элементы, необходимые для описания всех профессиональных занятий, перечисленных в *Словаре названий профессий (Dictionary of Occupational Titles)*. В добавление к этому, модель Кэмпбелла устанавливает три вида детерминант индивидуальных различий в выполнении работы, именно — декларативные знания, процедурные знания и навыки, мотивацию, а также их предпосылки (*antecedents*). Эта широкая и четко сформулированная теоретическая структура, по всей вероятности, найдет применение в самых разных исследованиях конструктов выполнения работы (*job performance*).

Использование тестов в сфере труда

Как следует из изложенного выше, правомерность использования тестов при принятии кадровых решений не может рассматриваться независимо от специфических целей, ситуаций и популяций, включенных в данный контекст.[1] Следует также отметить, что если мы и можем сгруппировать тесты по типам ради удобства рассмотрения, на практике демаркационные линии между знаниями, способностями (abilities), навыками и чертами личности не всегда столь отчетливы, как нам хотелось бы. Поэтому, возможно, было бы более продуктивно считать трудовую деятельность определяемой *потенциальными возможностями реагирования (response capabilities)*, как предлагали некоторые специалисты (Lubinski, & Dawis, 1992).[2] Кроме того, хотя валидность теста часто анализируют изолированно, сами тесты едва ли когда-нибудь применяются подобным образом. Большинство кадровых решений, принимаемых с учетом результатов тестирования, опираются на совокупную информацию, полученную с помощью одного или нескольких измерительных инструментов в сочетании с другими способами оценки — такими, как интервью или биографические опросники.[3] Помня обо всех этих предуведомлениях, перейдем к рассмотрению использования тестов в сфере профессиональной деятельности.

Роль академического интеллекта. «Интеллект» — широкий термин, имеющий множество определений. То, что включается в состав интеллекта, бесспорно, зависит от конкретной культуры, исторической эпохи и стадии возрастного развития (см. главы 11 и 12). Традиционные тесты интеллекта, напротив, охватывают более ограниченный кластер поддающихся раздельной оценке когнитивных умений и знаний, которые в большинстве своем оказались неплохими прогнозирующими параметрами

---

[1] Краткое рассмотрение многообразия переменных, влияющих на успешность взаимодействия «человек х контекст», дано в работе Sternberg (1994а).

[2] Сравнительный анализ понятий *skill*, *ability*, *aptitude* и *capacity* в контексте психологии труда можно найти в книге «Экспериментальная психология» под ред. С. С. Стивенса (гл. XXXVI. *Р. X. Сишор.* Работа и ее моторное исполнение. — С. 1011–1012). — *Примеч. науч. ред.*

[3] См. работу Guion (1991), где дан прекрасный обзор процесса оценки, отбора и расстановки кадров, включая данные о валидности и справедливости различных тестов и других типов предикторов.

уровня учебной и трудовой деятельности в современном индустриальном обществе. Вследствие того что в этот кластер входят, в основном, знания, умения и навыки, приобретенные в ходе обязательного в таком обществе школьного обучения, данный кластер часто описывают как академический интеллект или академические способности. Его содержание сводится, главным образом, к вербальному пониманию, количественному рассуждению и другим аспектам абстрактного мышления.

Хорошо известно, что показатели выполнения тестов академического интеллекта существенно коррелируют с объемом полученного образования. Почему бы тогда вместо применения этих тестов не ввести требования к объему образования, покрывающие квалификационные требования в области такого важного кластера знаний и навыков к кандидату на рабочее место? Однако на этом пути решения проблемы есть определенные трудности. Объем образования является косвенным показателем когнитивного статуса индивидуума на данный момент, и его корреляция с показателями тестов академического интеллекта далека от полной. Сам факт формального обучения еще не гарантирует одинакового усвоения учащимися всего того, чему учили в школе или колледже; более того, знания и умения, обычно формируемые в период формального обучения, *могут* приобретаться иными путями. Поэтому более справедливо тестировать знания и когнитивные умения конкретного человека, чем принимать или отказывать в приеме на работу на основе объема формального образования.

Среди серийно издаваемых инструментов есть несколько коротких тестов академического интеллекта, которые специально разрабатывались для применения в промышленности. Пример — Кадровый тест Вандерлика (Wonderlic Personnel Test, Inc., 1992). Появившийся в результате переработки одного из первых групповых тестов интеллекта — Самоприменяемых тестов умственных способностей Отиса (*Otis Self-Administering Tests of Mental Ability*), созданный Вандерликом инструмент представляет собой состоящий из 50 заданий 12-минутный тест. Он включает разные типы заданий с вербальным, числовым и пространственным содержанием, представленные в спиральном формате (*in a spiral-omnibus format*), и дает только один показатель. За несколько десятков лет использования теста Вандерлика, имеющего множество форм, по нему накоплены обширные нормативные данные для различных профессиональных групп, а его прогнозирующая сила в отношении успешности профессионального обучения и работы получила документальное подтверждение (что касается рецензий, см. Belcher, 1992; Schmidt, 1985; Schoenfeldt, 1985).

Интерес к потенциальной полезности общих тестов академического интеллекта для отбора персонала возобновился благодаря исследованиям, посвященным обобщению валидности (см., например, Hunter, 1986). В частности, были получены данные о том, что тесты вербального и числового рассуждения обладают в той или иной степени прогностической валидностью для широкого разнообразия профессиональных занятий. Вдобавок ко всему, эта валидность повышается для таких видов работы, которая требует более частого принятия решений и более полной обработки информации. Однако хотя тесты общих когнитивных способностей существенно облегчают предсказание уровня выполнения работы (*job performance*), особенно сложной, точность прогнозов можно повысить путем оценки дополнительных переменных. Главными среди таких переменных являются более специализированные знания и навыки, требующиеся для выполнения конкретных видов работы (в том числе психомоторные навыки и не выраженные в словах, или процедурные знания), а также некогнитивные переменные, такие как свойства темперамента и особенности аттитюдов (см., например,

Таблица 17–1

Факторы и комбинированные меры Батареи тестов общих способностей (GATB)

Факторы		
G. Общая способность к обучению (*General Learning Ability*)	S. Пространственная способность (*Spatial Aptitude*)	K. Моторная координация (*Motor Coordination*)
V. Вербальная способность (*Verbal Aptitude*)	P. Восприятие форм (*Form Perception*)	F. Быстрота и ловкость движений пальцев (*Finger Dexterity*)
N. Числовая способность (*Numerical Aptitude*)	Q. Восприятие канцелярских документов (*Clerical Perception*)	M. Быстрота и ловкость движений рук (*Manual Dexterity*)
Комбинированные меры		
Когнитивная = $G + V + N$	Перцептивная = $S + P + Q$	Психомоторная = $K + F + M$

Ackerman, 1992; J. P. Campbell, 1990 b; Carroll, 1992; Kanfer, Ackerman, Murtha, & Goff, 1995; Sternberg, Wagner, Williams, & Horvath, 1995). Многие из этих связей были вскрыты и подтверждены в крупномасштабных исследовательских проектах с применением классификационных батарей, используемых в вооруженных силах и в некоторых агентствах по найму гражданских госслужащих.

Батареи способностей для специальных программ.[1] Батарея тестов общих способностей (*General Aptitude Test Battery* [*GATB*]) была разработана Управлением размещения и регулирования рабочей силы США (*USES*) для использования консультантами по вопросам трудоустройства, работающими в организованных в каждом штате бюро по найму рабочей силы (U. S. Department of Labor, 1970). Кроме того, эту батарею могут получить некоммерческие организации, такие как средние школы, колледжи и тюрьмы. В настоящее время *GATB* включает 12 тестов, 4 из которых требуют для проведения простого оборудования, а остальные 8 относятся к тестам типа «бумага—карандаш». Проведение всей батареи тестов занимает примерно 2,5 ч. *GATB* дает показатели по 9 факторам и трем комбинированным мерам, получаемым из этих факторов, — все они перечислены в табл. 17–1.

Применение показателей *GATB* может строиться в соответствии с двумя различными подходами. Первый основан на использовании множественных критических показателей по наиболее важным способностям, необходимым для относительно однородных групп профессиональных занятий. Одним из механизмов реализации этого подхода является система Паттернов профессиональной пригодности (*Occupational Aptitude Pattern* [*OAP*]), разработанная в 1970-х гг. (U. S. Department of Labor, 1979, 1980). Паттерны профпригодности (*OAP*) были подготовлены для более 60 семейств профессий, охватывающих тысячи конкретных специальностей. Для каждой группы профессий были вычислены критические показатели высокого, среднего и низкого уровней пригодности по соответствующим способностям, которые могут использо-

[1] Предварительные сведения к этому разделу см. в главах 10 и 11.

ваться в профконсультировании.[1] Второй подход к использованию *GATB* сложился в результате применения процедур обобщения валидности (*VG*) к данным из более 500 ранее проведенных *USES* исследований валидности. Эта процедура, которую стали называть *VG-GATB*, использует оценки валидности, основанные на соответствующих комбинациях показателей для всех специальностей, входящих в каждое семейство профессий (U. S. Department of Labor, 1983a, 1983c, 1983d). Предсказания строятся на основе трех комбинированных мер — когнитивной, перцептивной и психомоторной, — выводимых из показателей по первичным факторам. Из этих трех мер когнитивная комбинированная оценка дает наибольшие коэффициенты валидности для большинства специальностей, но психомоторная комбинированная мера может улучшать прогноз в тех случаях, когда снижается уровень сложности выполняемых работ (Hunter, & Hunter, 1984).

Благодаря возможностям такой мощной организации, как *USES*, был накоплен обширный корпус данных по *GATB*, большей частью подтверждающих исключительную надежность и существенную прогностическую валидность этой батареи (см. рецензии B. Bolton, 1994; Kirnan, & Geisinger, 1986). Однако практика субгруппового нормирования, введенная в 1980-х гг. вместе с *VG-GATB* для обеспечения сопоставимости относительного количества направлений на работу, полученных белыми, черными и испаноязычными кандидатами, привела к горячим политическим дебатам по поводу справедливости тестирования перед направлением на работу (Hartigan, & Wigdor, 1989). Споры достигли кульминации с принятием закона о гражданских правах от 1991 г. (*Civil Rights Act of 1991* — P.L. 102–166), запрещающего установление норм для подгрупп. Этот законодательный акт повлиял на использование Батареи тестов общих способностей и сделал ее будущее неопределенным (L. S. Gottfredson, 1994; Wigdor, & Sackett, 1993 — см. также главу 18). Тем не менее исследовательская программа, включающая разработку двух новых форм и экспериментального образца компьютеризованной адаптивной версии *GATB*, продолжает выполняться.

Другим важным средством отбора и распределения персонала является Батарея профессиональной пригодности вооруженных сил США (*Armed Services Vocational Aptitude Battery* [*ASVAB*]), разработанная совместно для использования во всех родах войск США (Bayroff, & Fuchs, 1970). Эта батарея проводится с учащимися средних школ, проявляющим интерес к военным специальностям, а также со всеми, кто обратился с заявлением о желании поступить на армейскую службу. Современные формы *ASVAB* включают 10 субтестов, перечисленных в табл. 17–2.[2] Квалификационный тест вооруженных сил (*Armed Forces Qualification Test* [*AFQT*]) есть не что иное, как комбинированный показатель, используемый всеми армейскими службами в качестве меры общей обучаемости (*general trainability*) при предварительном отборе по-

---

[1] Группы Батареи тестов специальных способностей (*Special Aptitude Test Battery* [*SATB*]) обеспечивают альтернативный механизм для использования критических показателей наряду с *GATB*. Стратегия мультикритериального отсева более подробно обсуждается в главе 6. В области профотбора, в отличие от профконсультирования, наиболее подходящим применением критических показателей считается предварительный отсев кандидатов по одному или более необходимым для конкретной работы навыкам.

[2] Имеется также компьютеризованная адаптивная версия *ASVAB* (*CAT-ASVAB*), разработка которой велась с 1979 г. Теперь она введена в эксплуатацию и используется для обработки входных данных по некоторым военным специальностям (T. L. Russell, Reynolds, & Campbell, 1994). Описание *CAT-ASVAB* и особенностей ее разработки см. в работе Wiskoff, & Schratz (1989).

Таблица 17-2
Батарея профессиональной пригодности вооруженных сил США (*ASVAB*)

Субтесты *ASVAB*	
Общие естественнонаучные знания (*General Science* [*GS*])	Арифметическое рассуждение (*Arithmetic Reasoning* [*AR*])*
Знание слов (*Word Knowledge* [*WK*])*	Математические знания (*Mathematics Knowledge* [*MK*])*
Понимание параграфов инструкций (*Paragraph Comprehension* [*PC*])*	Понимание механических закономерностей (*Mechanical Comprehension* [*MC*])
Осведомленность в электронике (*Electronics Information* [*EI*])	Компетентность в автотехнике (*Auto and Shop Information* [*AS*])
Скорость кодирования (*Coding Speed* [*CS*])**	Числовые операции (*Numerical Operations* [*NO*]) **

* — Составная часть комбинированного показателя AFQT. ** — Скоростной тест.

тенциальных новобранцев. В добавление к этому, каждая из служб объединяет субтесты таким образом, чтобы создавать комбинированные показатели для отбора и распределения персонала в соответствии с собственными нуждами. Например, для боевых подразделений комбинированный показатель образуется путем суммирования показателей следующих субтестов: *AR* + *CS* + *AS* + *MC*. Стандартные показатели для *ASVAB* основываются на нормах, полученных при обследовании репрезентативной выборки американских юношей (U. S. Department of Defense, 1982). Факторная структура батареи изучена довольно подробно. Типичные результаты в этом случае свидетельствуют о наличии одного общего фактора, объясняющего 60 % полной дисперсии *ASVAB*, и четырех групповых факторов, которые неоднократно выделялись разными исследователями (J. R. Welsh, Watson, & Ree, 1990). Эти четыре фактора вместе с субтестами, по которым они имеют наибольшие нагрузки, таковы: 1) Вербальный (*WK* и *PC*), 2) Скоростной (*NO* и *CS*), 3) Количественный (*AR* и *MK*) и 4) Технический (*AS*, *MC* и *EI*).

Валидность субтестов *ASVAB* и получаемых на их основе комбинированных показателей исследовалась относительно широкого множества критериев учебной и профессиональной деятельности. Нетрудно догадаться, что коэффициенты валидности существенно различаются в зависимости от типа и количества используемых критериев. В общем, коэффициенты валидности выше для критериев «исполнения» (*can-do*), таких как общий срок службы в армии и техническая квалификация, чем для критериев «намерения» (*will-do*), — наподобие служебного рвения, лидерства и личной дисциплины. Как и следовало ожидать, первые оцениваются при помощи мер профессиональных знаний и «реальных дел» (*hands-on*), тогда как последние — с помощью самооценок и рейтингов военнослужащих, получаемых от их непосредственных начальников и сослуживцев (McHenry, Hough, Toquam, Hanson, & Ashworth, 1990; T. L. Russell et al., 1994).

Самое обширное исследование *ASVAB* было начато министерством обороны США в 1980 г. как часть Проекта объединенной комиссии по разработке стандартов измерения выполнения работы и зачисления на военную службу (*Joint-Service Job Performance Measurement/Enlistment Standards* [*JPM*] *Project* — Wigdor, & Green, 1991a, 1991b). Цель этого крупного проекта — разработать устойчивые («робастные») меры выполнения работы для начального уровня военных специальностей и на их основе устано-

вить значимые и обоснованные стандарты для зачисления добровольцев на службу во все рода войск. Первая фаза реализации проекта *JPM* дала все основания считать *ASVAB* хорошим предиктором воспроизводимых с высокой точностью, основанных на выполнении реальной работы признаков профессионального мастерства.[1] Кроме того, были получены данные, подтверждающие дифференциальную валидность комбинированных показателей *ASVAB* в отношении разных специальностей. Однако величина различий в средних показателях между черными и не принадлежащими к меньшинствам испытуемыми оказалась существенно больше по *AFQT* и бланковым тестам профессиональных знаний, чем по результатам выборочных проверок реальной деятельности (*hands-on job sample tests*). Следовательно, существует возможность того, что некоторые из мер *ASVAB* могут переоценивать величину фактических групповых различий в выполнении работы. Если это предположение подтвердится, данная ситуация будет иметь сходство с некоторыми результатами, полученными при применении *GATB* (Hartigan, & Wigdor, 1989). К тому же, коэффициенты валидности *ASVAB* относительно различных критериев оказались достаточно умеренными, чтобы оправдать поиск дополнительных предикторов. Вторая фаза проекта *JPM* посвящена предварительному исследованию моделей стандартов для зачисления на военную службу, которые, предположительно, должны повысить общую полезность решений по отбору и классификации персонала, — как с точки зрения затрат, так и с позиции уровней выполнения профессиональных функций.

Проект отбора и распределения специалистов сухопутных войск США (Проект А) охватывает другой важный сегмент исследований *ASVAB* и новых предикторов выполнения профессиональных функций военными специалистами. Проект А возник как позиционный ответ армии на полномочия проекта *JPM* и, по-видимому, действительно является самым крупным и дорогостоящим из всех когда-либо осуществлявшихся проектов исследований отбора персонала (Schmidt et al., 1992). В дополнение к своему вкладу в построение теории профессиональной деятельности (обсуждавшейся выше в этой главе), Проект А связан с разработкой и оценкой многих новых предикторов, которые выходят за пределы традиционных когнитивных функций *ASVAB*. Создаваемая в рамках армейского проекта батарея включает компьютеризованные перцептивные и психомоторные тесты; специально сконструированные опросники для оценки параметров личности, темперамента и жизненного опыта, а также инвентари интересов (McHenry et al., 1990; N. G. Peterson et al., 1990). Кроме того, масштабность и лонгитюдный характер Проекта А обеспечили беспрецедентную возможность изучения временных изменений валидности (J. P. Campbell, 1990b).[2]

Тесты специальных способностей. Еще до создания комплексных батарей способностей многие специалисты сознавали, что тесты интеллекта охватывают далеко не все человеческие способности, и вскоре были предприняты попытки заполнить основные пробелы с помощью тестов специальных способностей, предназначенных для измерения более конкретных и практических способностей, наподобие механических.

---

[1] Теоретическая и методологическая работа, проведенная в ходе разработки прямых мер выполнения профессиональной деятельности, является ценным вкладом проекта *JPM* в область измерений, включая измерения в образовании.

[2] Заключительный отчет по Проекту А в виде коллективной монографии под совместной редакцией Дж. Кэмпбелла (J. P. Campbell) и Д. Нэппа (D. Knapp) планируется опубликовать в конце 1990-х гг.

Требования профотбора и профконсультирования тоже стимулировали разработку средств измерения пространственных, канцелярских, музыкальных и художественных способностей. С другой стороны, тесты зрения, слуха, мышечной работы (*muscular performance*) и ловкости движений (*motor dexterity*) широко использовались при отборе и классификации персонала для промышленных и военных целей.[1]

Нужно добавить несколько слов о самом понятии *специальные способности (special aptitudes)*. Термин возник в то время, когда главное место в тестировании отводилось измерению общего интеллекта. Технические, музыкальные и другие специальные способности рассматривались, таким образом, лишь как дополнение к *IQ* при характеристике конкретного человека. Однако с появлением факторного анализа пришло постепенное осознание того, что сам интеллект состоит из ряда относительно независимых способностей, таких как вербальное понимание, числовое рассуждение, оперирование пространственными образами и т. д. Более того, ряд способностей, традиционно рассматриваемых как специальные, например механические и канцелярские, теперь включаются в некоторые комплексные батареи способностей.

Какова же в таком случае роль тестов специальных способностей? Во-первых, такие области, как зрение, слух, ловкость движений (сноровка) и художественное дарование, редко включаются в комплексные батареи способностей. Ситуации, требующие тестирования именно этих областей, являются слишком специализированными, чтобы было оправдано включение подобных тестов в стандартные батареи. Тем не менее тесты специальных способностей используются и в тех областях, которые охватываются комплексными батареями способностей, например, таких как канцелярские и механические способности. В некоторых программах тестирования тесты академического интеллекта объединяют со специально отобранными тестами других, соответствующих целям конкретной программы, способностей. Одна из причин такой практики — получить доступ к широким нормативным и валидационным данным, пригодным для некоторых широко используемых тестов специальных способностей. Другой причиной, бесспорно, является гибкость такого подхода, проявляющаяся не только в выборе релевантных способностей, но и в полноте измерения каждой отдельной способности для конкретных целей.

Довольно много тестов было придумано для измерения скорости и координации движений, а также других *психомоторных навыков (psychomotor skills)*. Большинство тестов оценивает мануальную ловкость (*manual dexterity*), и лишь несколько — движения ног или ступней, необходимые при выполнении конкретных видов работы. Некоторые тесты измеряют комбинацию моторики с перцептивными, пространственными и механическими способностями. В основном, эти тесты применялись для отбора персонала в промышленности и армии. Психомоторные тесты, как правило, относятся к типу аппаратурных тестов, хотя несколько их адаптаций типа «бумага—карандаш» были разработаны для группового проведения. Примером выпускаемого в настоящее время инструмента для оценки нескольких простых манипулятивных навыков может служить Тест ловкости оперирования мелкими деталями Кроуфорда (*Crawford Small Parts Dexterity Test* [*CSPDT*] — Crawford, & Crawford, 1981), общий вид которого показан на рис. 17–2. В первой части этого теста испытуемый должен, пользуясь пинцетом, вставить штифты в точно соответствующие их диаметру отверстия, а затем

[1] Обзор методик отбора кадров для специальностей, предъявляющих требования к физическим данным, имеется в работе J. C. Hogan (1992).

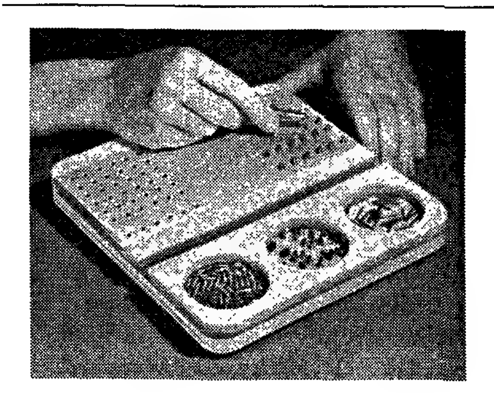
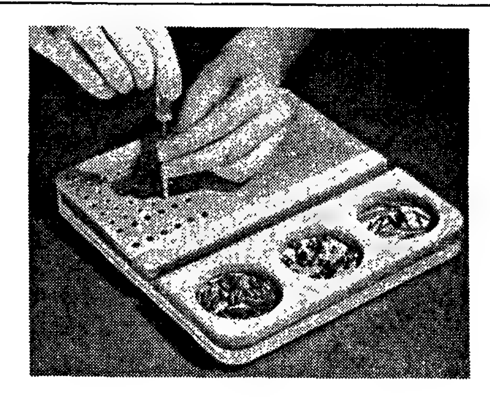



Рис. 17–2. Тест ловкости оперирования мелкими деталями Кроуфорда.
(С любезного разрешения Психологической корпорации)

на каждый штифт надеть маленькую узкую втулку. Во второй части маленькие винтики вставляются в имеющие резьбу отверстия и завинчиваются с помощью отвертки. Показателем является время, затраченное на выполнение каждой части теста.

Что можно сказать об эффективности психомоторных тестов в целом? Наиболее важным моментом, который необходимо отметить при оценивании таких тестов, является высокая степень *специфичности* измеряемых с их помощью моторных функций. Корреляционный и факторный анализ большого количества моторных тестов не выявил широких групповых факторов, наподобие тех, что обнаружены для интеллектуальных функций (Fleishman, 1975; Fleishman, & Quaintance, 1984, chap. 12). При рассмотрении валидности психомоторных тестов нужно различать сложные моторные тесты, имеющие тесное сходство с конкретной критериальной деятельностью, и тесты простых моторных функций, предназначенные для общего использования. Первые относятся к разрабатываемым по индивидуальному заказу и, как правило, теперь уже компьютеризованным тестам, которые воспроизводят сочетание моторных способностей, предполагаемых критерием, и демонстрируют достаточную валидность. Так, например, было обнаружено, что некоторые тесты ВВС США улучшают предсказание уровня пилотирования после окончания летных курсов (см., например, R. H. Cox, 1989; Kantor, & Carretta, 1988). Но для большинства целей использовать подобные тесты практически невозможно, так как пришлось бы придумывать очень большое число тестов, согласующихся с многообразными критериями. Что касается серийно выпускаемых и доступных для приобретения моторных тестов, то измеряемые ими функции весьма просты, а их валидность относительно большинства критериев не слишком высока. Вот почему такие тесты предпочтительней применять в составе используемой для отбора батареи, а не в качестве отдельных предикторов.

Тесты *механических способностей (mechanical aptitude)* охватывают целый ряд функций. Психомоторные факторы входят в некоторые тесты этой категории либо потому, что выполнение теста требует умения быстро манипулировать с тестовым материалом, либо потому, что специальные субтесты для измерения ловкости движений включаются в тот или иной бланковый тест. Перцептивные и пространственные способности также играют важную роль во многих таких тестах. И, наконец, в ряде тестов механических способностей господствующее положение отводится осведомленности в области механики и умению рассуждать на этом материале.

Важно отдавать себе отчет в разнообразии функций, объединенных под рубрикой механических способностей, так как каждая функция может быть по-разному связана с другими переменными. Например, тесты на осведомленность в области механики гораздо больше зависят от прошлого опыта обращения с механическими устройствами, чем абстрактные пространственные или перцептивные тесты. Аналогично этому, половые различия могут приобретать прямо противоположный характер при переходе от одной функции к другой. Так, в тестах на мануальную ловкость и перцептивное различение женщины обычно превосходят мужчин; в абстрактных пространственных тестах обычно выявляются небольшие, но значимые средние различия в пользу мужчин, а в тестах на механическое рассуждение и осведомленность в области механики мужчины демонстрируют уже заметное преимущество (Anastasi, 1981c; Hedges, & Nowell, 1995).

Среди способностей, включаемых в состав всех комплексных батарей способностей, оказывается пространственная способность (*spatial aptitude*) или, иначе говоря, способность оперирования пространственными образами. Именно эта способность измеряется тестом «Пространственные отношения» из батареи *DAT* (см. главу 10), и она же, как оказалось, имеет высокую нагрузку во многих тестах действия и неязыковых тестах общего интеллекта. Одним из самых простых средств измерения этой способности является Пересмотренный миннесотский бланковый тест «Доска форм» (*Revised Minnesota Paper Form Board Test* [*RMPFBT*] — Likert, & Quasha, 1995).

Другой важный тип тестов механических способностей касается осведомленности в области механики, механических рассуждений и понимания механических закономерностей. Хотя эти тесты требуют некоторого знакомства с механическими орудиями труда и законами механики, они предполагают у тестируемых лишь такой объем технических знаний, который можно приобрести из повседневного опыта жизни в современных промышленно развитых странах. Некоторые из ранних тестов в этой области требовали от испытуемого собрать из предлагавшихся деталей простые, широко используемые механические устройства. Когда проводится неспециализированное тестирование, вместо таких тестов теперь широко применяются групповые тесты типа «бумага-карандаш». Известным примером этого типа тестов является Тест понимания механических закономерностей Беннетта (*Bennett Mechanical Comprehension Test* [*BMCT*] — G. K. Bennett, 1994). Используемые в этом тесте короткие вопросы по картинкам требуют для ответа на них понимания законов механики, применяемых в самых разных ситуациях повседневной жизни. Два ознакомительных задания из этого теста воспроизведены на рис. 17–3.

Тесты, предназначенные для измерения *канцелярских способностей* (*clerical aptitudes*), характеризуются общим акцентом на скорости и точности восприятия. Типичный пример — Миннесотский канцелярский тест (*Minnesota Clerical Test* [*MCT*] — Andrew, Paterson, & Longstaff, 1979), состоящий из двух раздельно оцениваемых по времени выполнения субтестов: Сравнение чисел (*Number Comparison*) и Сравнение названий (*Name Comparison*). В первом субтесте испытуемому даются 200 пар чисел, каждая из которых содержит от 3 до 12 цифр. Если числа, составляющие пару, одинаковы, то испытуемый ставит между ними галочку. Во втором субтесте задача та же самая, но вместо чисел испытуемому предъявляются слова.

Такие относительно однородные тесты, как Миннесотский канцелярский тест, измеряют только один аспект работы служащего. Конторские же работы связаны с выполнением многочисленных функций. Более того, число и конкретное сочетание обя-

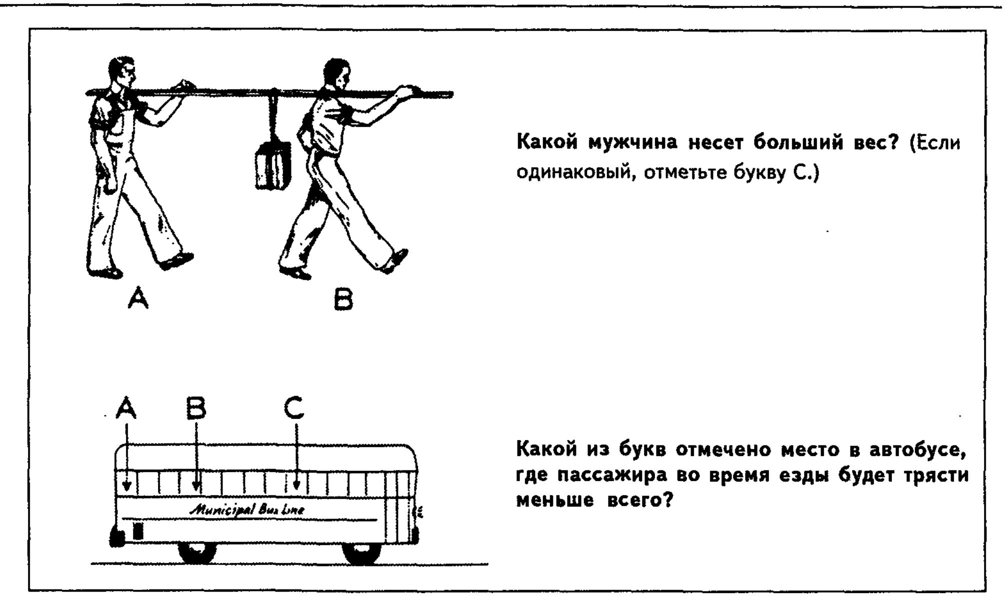


Рис. 17–3. Пробные задания из Теста понимания механических закономерностей Беннетта. Ответы записываются на отдельном бланке для ответов

(Воспроизводится с разрешения. Copyright © 1967–1970, 1994 by The Psychological Corporation. All rights reserved)

занностей канцелярского работника сильно зависят от типа и уровня выполняемой им работы. Однако несмотря на такое разнообразие занятий, анализ содержания работы клерка показывает, что относительно большая часть времени тратится на выполнение заданий, требующих скорости и точности в восприятии деталей.

Разумеется, перцептивная скорость и точность необходимы не только в конторской работе. Инспекторам, контролерам, упаковщикам и многим другим работникам промышленных предприятий явно нужна эта способность, хотя большинство таких видов работы теперь выполняют электронные сканеры.

Некоторые тесты канцелярских способностей предполагают сочетание перцептивной скорости и точности с другими функциями, необходимыми для конторской работы. К числу используемых инструментов такого рода относятся тесты выборочных образцов реальной работы (*job-sample tests*) для оценки таких видов деятельности, как упорядочение материалов по алфавиту, классификация, кодирование и т. п. В добавление к этому, могут включаться некоторые средства измерения вербальных и числовых способностей, которые служат вместо тестов общего интеллекта. Другие тесты канцелярских способностей предполагают оценку таких навыков канцелярской работы, как знание деловой терминологии, владение деловой информацией, умение писать и говорить без ошибок. Некоторые канцелярские тесты более правильно классифицировать как тесты достижений или тесты выборочных образцов работы, так как они измеряют навыки, приобретенные после прохождения специального обучения. Примеры включают тесты стенографических навыков и навыков машинописи, так же как и тесты более современных навыков ввода данных и поиска информации в базах данных, наподобие Теста навыков работы с видеотерминалом Объединения научных исследований (*CRT Skills Test by Science Research Associates* — SRA, 1990).

Быстрый рост использования компьютеров в канцелярской работе привел к тому, что был выпущен целый ряд тестов для оценки *способностей, связанных с применением вычислительных машин (computer-related aptitudes).* Многие из них предназначались для консультирования или отбора потенциальных операторов-учеников. Примерами могут служить Тесты теоретических и практических знаний в области вычислительной техники (*Computer Literacy and Computer Science Tests*) и Батарея способностей программиста ЭВМ (*Computer Programmer Aptitude Battery*).[1] Хотя эти тесты представляли собой передовые приложения психометрии в области оценки персонала в то время, когда они разрабатывались, — в период между 1960-ми и началом 1980-х гг., — технический прогресс идет настолько быстрыми темпами, что некоторые из этих инструментов уже полностью устарели. Постепенно появляются новые инструменты для тестирования пригодности к обучению и компетентности пользователей конкретного программного обеспечения, скажем, *dBASE, WordPerfect* и *Lotus* 1-2-3.[2]

Наблюдающийся в последнее время прогресс в признании важности когнитивных измерений (*dimensions*) межличностного и личностного функционирования (см., например, H. Gardner, 1983; Salovey, & Mayer, 1990), по-видимому, стимулирует разработку стандартизованных объективных инструментов для оценки *социальных и эмоциональных аспектов интеллекта* в контексте трудовой деятельности. Раньше таких инструментов практически не было, так как адекватность межличностного и личностного функционирования в рабочей обстановке традиционно измерялась либо тестами личности, либо с помощью интервью и методик оценки в центрах. Тест бригадной работы-*KSA* (*Teamwork-KSA Test*) — недавно выпущенная бланковая методика для предсказания эффективности коллективной работы — является одним из первых инструментов в составе этой новой категории тестов. Опираясь на анализ литературы по групповой работе в организациях, авторы данного теста попытались создать стандартизованное средство для измерения знаний, навыков и способностей работников в таких областях, как межличностное взаимодействие и организация собственной деятельности (*self-management*). Задания с множественным выбором предлагают тестируемым гипотетические вопросы, касающиеся разрешения конфликтов, коммуникации и совместного решения проблем, а также постановки целей, планирования и других навыков самоорганизации (M. J. Stevens, & Campion, 1994). Дальнейшая работа и экспериментирование с этим и другими аналогичными инструментами, безусловно, будут продолжены.

Тестирование личности работников

В середине 1980-х гг. Бернардин и Боунас (Bernardin, & Bownas, 1985) отметили, что, в то время как методики оценки личности работников (включая такие ненаучные методы, как анализ почерка) широко использовались в промышленности, научное сообщество на протяжении почти двух десятков лет, по существу, игнорировало эту

[1] Что касается критических разборов этих инструментов, см. Mahurin (1992), Marco (1992), Schafer (1992).

[2] Эти и некоторые другие инструменты такого типа можно заказать в SRA Product Group of McGraw-Hill / London House (см. приложение Б). (Перечисленные авторами программные продукты устарели уже на момент выхода оригинала этой книги, а в настоящее время практически вышли из употребления. — *Примеч. науч. ред.*)

тему.[1] С того времени произошло невиданное оживление исследований в этой области, подсказываемых методологическими и теоретическими достижениями. С методологической точки зрения, применение метаанализа и использование методов причинного моделирования (см. главу 5) стимулировали изучение тех некогнитивных черт, которые могут влиять на выполнение работы. Метаанализ применялся, главным образом, для исследования валидности и полезности конструктов личности в разной обстановке. Методы путевого анализа и моделирования структурными уравнениями (см. главу 5) использовались для изучения взаимосвязей между предикторами и выявления характеристик и состояний, приводящих к различным уровням выполнения работы. Изучается и то, в какой степени некоторые критические переменные, наподобие уровня способности и автономности работы, ослабляют взаимосвязь между личностью и деятельностью. В общем, цели этих исследований идут дальше прогнозирования, к пониманию конструктов и процессов, обусловливающих те значительные вариации, которые имеют место в выполнении работы (Barrick, & Mount, 1991, 1993; Hough, Eaton, Dunnette, Kamp, & McCloy, 1990; Matthews, Jones, & Chamberlain, 1992; McHenry et al., 1990; Schmidt & Hunter, 1992; Tett, Jackson, & Rothstein, 1991).

Более изощренные в методологическом отношении исследования последних двух десятилетий оказали существенную поддержку применению тщательно сконструированных личностных тестов при принятии кадровых решений в разнообразных контекстах. Важный и пока еще недостаточно исследованный вопрос касается определения наиболее релевантных измерений (*dimensions*) личности по отношению к конкретным специальностям или семействам специальностей. Например, в то время как одни виды работы могут требовать высокой степени общительности, другие требуют от работника прямо противоположного качества. Даже такая характеристика, как сознательность (*conscientiousness*), которая на первый взгляд может показаться желательной для всякой работы, при более глубоком рассмотрении вдруг оказывается связанной с характеристиками, мешающими определенным занятиям, — таким, например, как творческие искания, — по меньшей мере в некоторых отношениях. В результате учета подобных соображений в настоящее время начинает уделяться внимание установлению требований конкретной работы к особенностям темперамента и межличностным характеристикам работников (R. Hogan, Hogan, & Roberts, 1996; Landy et al., 1994).

Значительная часть текущих исследований сосредоточена на выяснении полезности «Пятифакторной модели» структуры личности — в ее различных формах — для предсказания критериев выполнения работы (см. главу 13 и R. Hogan, 1991). Популярность этой модели в исследованиях отбора персонала не случайна. В конце концов, эти пять (± 2) факторов были извлечены в результате обработки обширных массивов данных описания личности и представляют собой модель для характеристики нормальных личностей, в отношении которой существует довольно высокая степень согласия среди специалистов. Каждая из областей, охватываемых данной моделью, широко применима ко всем видам повседневного поведения, включая выполнение профессиональных обязанностей. Например, измерение «Эмоциональная устойчивость»

[1] См. работу Anastasi (1985e), где дан исторический обзор тестирования личности в промышленности и рассматриваются основные методологические и практические проблемы, связанные с этой областью тестирования. В работе Kanfer et al. (1995) проведен анализ прогресса в приложении понятий различных теорий личности и интеллекта к области промышленной и организационной психологии.

(*Emotional Stability*)[1] является существенным в работе, связанной с быстрым принятием решений в условиях психического напряжения, такой как охрана правопорядка, пилотирование самолета, вождение большегрузных автомобилей и скорая медицинская помощь. С другой стороны, «Уживчивость» (*Agreeableness*) является совершенно необходимым качеством для любой работы, которая предполагает широкие контакты с людьми. Впрочем, неудивительно, что именно фактор «Сознательность» (*Conscientiousness*) выявился в целом ряде метааналитических исследований как наиболее универсальный и существенный личностный предиктор выполнения работы (Barrick, & Mount, 1991, 1993; Schmidt, & Hunter, 1992). Тем не менее пока еще далеко до полного согласия как по поводу определения конструкта «Сознательность», так и в отношении признания его первенства (см., например, Loevinger, 1994; Tett et al., 1991).

Тесты честности. Применение тестов честности, или правдивости, при принятии решений о приеме на работу приобрело особую актуальность после принятия закона о защите наемных работников от проверки на полиграфе от 1988 г. (*Employee Polygraph Protection Act of 1988* [*P. L.* 100–347]), который запретил использование детекторов лжи при отборе персонала, за исключением особых условий, как при приеме служащих в правительственные организации.[2] Одним из очевидных результатов принятия этого закона стало быстрое распространение бланковых средств измерения надежности работников, получивших широкую известность как «тесты честности» (*integrity tests*). Эти средства, которые обычно представляют собой набор вопросов, нацеленных на выяснение аттитюдов кандидата в отношении воровства и других видов незаконного поведения, а также его возможной замешанности в прошлом в такого рода поведении, сами быстро оказались предметом тщательной проверки в психологии и за ее пределами.[3] В одном из наиболее полных исследований тестов честности Оунс, Висвешваран и Шмидт (Ones, Viswesvaran, & Schmidt, 1993) провели метаанализ, базирующийся на 665 коэффициентах валидности. Их исследование позволило подсчитать, что средняя рабочая валидность (*operational validity*) тестов честности в отношении предсказания даваемых начальниками оценок выполнения работы равна 0,41 и свидетельствует о ценности этих тестов для прогнозирования поведения, вредящего работе.

Несмотря на весьма внушительную степень поддержки применения тестов честности, правда, в рамках определенных границ, их использование в отборе персонала — на этапе принятия решений о приеме на работу или отказе — остается проблематичным. В настоящее время специалистов заботит возможная подверженность тестов честно-

---

[1] Рассматриваемая как свойство, противоположное нейротизму (эмоциональной неустойчивости). — *Примеч. науч. ред.*

[2] Книга Ликкена (Lykken, 1981) — одна из первых попыток критического анализа проблемы распознавания лжи и применения для этой цели психофизиологических методов, предпринятая психологом. Более краткое и современное рассмотрение проблем применения полиграфов можно найти в работах DePaulo (1994), Honts (1994), Kircher, & Raskin (1992), Lykken (1992), Saxe (1994).

[3] См., например, отчет специальной комиссии, сформированной Американской психологической ассоциацией для углубленного изучения данного вопроса (L. R. Goldberg, Grenier, Guion, Sechrest, & Wing, 1991). Более подробную информацию содержит объемный труд К. Мёрфи (K. R. Murphy, 1993), посвященный общей проблеме честности на рабочем месте. Краткое, но содержательное обсуждение проблем тестирования честности при отборе персонала, можно найти в работах Camara, & Schneider (1994), Sackett (1994) и в разделе «Критика» июньского номера журнала *American Psychologist* за 1995 г.

сти влиянию натаскивания на «правильные» ответы и сознательным попыткам кандидатов создать о себе благоприятное мнение, а также относительной неэффективностью этих тестов в предсказании конкретных форм вредящего работе поведения, например воровства (см., например, Alliger, Lilienfeld, & Mitchell, 1996; Camara & Schneider, 1995).

Лидерство. Отбор эффективных лидеров[1] ставит еще одну важную проблему в области кадровых решений. Лидерство — одно из наиболее востребованных качеств работников, потому что оно включает в себя способность склонить других к работе ради общего дела. Эффективное лидерство может значительно улучшить функционирование организации, тогда как неэффективное лидерство может привести к гибельным последствиям. Поэтому вряд ли стоит удивляться, что многие типы средств измерения когнитивных и личностных переменных, а также методики оценки в центрах, интервью, моделирования и сбора биографических данных применяются для подбора высшего исполнительного руководства. Чаще всего процесс оценки, входящий составной частью в отбор руководителей высшего звена, предполагает использование многих методов и длится несколько часов или даже дней. Такие процедуры очень напоминают клиническое тестирование и обычно являются дорогостоящими.

Очевидно, что в большинстве случаев средства, выделяемые на тщательный отбор лиц, включаемых в кадровый резерв организации и постепенно продвигаемых по служебной лестнице до ключевых руководящих должностей, оказываются весьма скудными по сравнению с потенциальным воздействием этих решений. Возможно, что именно вследствие такого положения дел управленческая некомпетентность достигла невероятных масштабов (R. Hogan, Curphy, & Hogan, 1994). Хотя интерес к изучению разных видов лидерства имеет давнюю историю в прикладной психологии, — о чем свидетельствует обширная литература по данной теме, — многое еще предстоит сделать в области выявления и уточнения эмпирических коррелятов эффективного лидерства. Более полную информацию о средствах измерения лидерства, а также о теоретических и методологических успехах в изучении лидерства в организациях можно найти в работах Bass (1990), Clark, & Clark (1990), Yukl, & Van Fleet (1992).

Инструменты. За исключением методик интенсивной оценки, применяемых при отборе высшего исполнительного руководства крупных организаций и некоторых категорий специалистов, — таких как разведчики, астронавты и т. п., — подавляющему большинству кандидатов , которым в ходе собеседования предлагают пройти тестирование личности, дают для заполнения опросники наподобие тех, что обсуждались в главе 13. Такие опросники, как *MMPI*, нацеленные на обнаружение психопатологии, все еще применяются для целей предварительного отсеивания в области некоторых профессий, связанных с государственной или военной тайной, но беспокойство общественности по поводу вмешательства в личную жизнь граждан и злоупотребления результатами тестирования привело к преобладающему использованию инструментария, специально разработанного для оценки личности нормальных людей (см. главу 18). Фактически, несколько относительно старых многомерных опросников, построенных в форме стандартизованных самоотчетов и изначально предназначавшихся для нормальных испытуемых, — например, Калифорнийский психологический

[1] Разумеется, речь идет о формальных лидерах. — *Примеч. науч. ред.*

опросник (*CPI*), — были недавно пересмотрены с целью сделать их более подходящими для применения в сфере труда. Тем временем был разработан ряд более новых инструментов, наподобие Личностного опросника Хогана (*Hogan Personality Inventory* [HPI] — R. Hogan, & Hogan, 1992), предназначенных в первую очередь для использования в профконсультировании и профотборе. Другие тесты разрабатываются для решения более узких задач; примеры включают *PDI* опросник службы работы с покупателями (*PDI Customer Service Inventory*) и *PDI* опросник службы по трудоустройству (*PDI Employment Inventory*). Оба являются инструментами для предварительного отбора наемных служащих, от которых работа потребует в любых ситуациях внушать своими действиями доверие и проявлять стойкость (см. Paajanen, Hansen, & McLellan, 1993).

Заключительные замечания. Достижения современных технологий преобразуют характер труда в направлениях, далеко выходящих за пределы изменений, вызванных промышленной революцией. Быстрый темп этих перемен ставит перед специалистами в области психологии труда проблемы огромной важности, а прогресс в методологии, ставший возможным благодаря новым технологиям, предлагает невиданные ранее возможности для решения таких проблем.[1] Теперь, когда появилась возможность проводить масштабные и методологически изощренные исследования, позволяющие получить, в известной степени, окончательные ответы на вопросы, занимавшие специалистов в течение долгого времени, начинают изменяться сами эти вопросы. Между тем влияние компьютерных технологий на тестирование персонала хотя и заметно, но пока еще слабо и носит скорее общий характер (Bartram, 1993; Burke, 1993; Schoenfeldt, & Mendoza, 1991). Некоторые из наиболее интересных приложений современных технологий к тестированию — такие, как интерактивные мультимедийные тесты, — только начинают развиваться (Desmarais, Masi, Olson, Barbera, & Dyer, 1994; Drasgow et al., 1996). В этой области гораздо больше прогнозов и ожиданий, чем реальных достижений. В целом, однако, в настоящее время существует больше чем когда-либо благоприятных возможностей для продуманного применения психологических принципов и методов в такой сфере, как улучшение использования трудовых ресурсов (Bray et al., 1991; Gatewood, & Feild, 1993).

Использование тестов в клинической психологии и психологическом консультировании

Психологи-клиницисты и психологи-консультанты применяют достаточно разнообразные тесты, и большинство из уже рассмотренных нами типов тестов можно встретить в их практике. Стало традицией проводить периодические опросы выборок клинических психологов и психологов-консультантов, различающихся своей теоретической ориентацией и местом работы, в отношении используемых ими тестов (Archer, Maruish, Imhof, & Piotrowski, 1991; Lubin, Larsen, Matarazzo, 1984; Piotrowski, & Keller, 1992; Watkins et al., 1994; Watkins, Campbell, Nieberding, & Hallmark, 1995). Их резуль-

---

[1] Обсуждение преобразований, происходящих в настоящее время на рабочих местах, и рассмотрение тех последствий, которые такие преобразования могут иметь для работы с кадрами, можно найти в работе Landy, Shankster-Cawley, & Kohler Moran (1995).

таты свидетельствуют о том, что весьма часто используются такие индивидуальные тесты интеллекта, как шкалы Векслера, и комплексные батареи способностей, наподобие *DAT* (см. главу 10). Многие из личностных тестов, рассмотренных в главах 13–16, занимают заметное место в методическом репертуаре клиницистов и консультантов. К некоторым диагностическим образовательным тестам обращаются при неспособности отдельных учащихся справиться с требованиями учебного плана или при возникновении других, связанных со школой, проблем. В добавление к этому, клиницисты и консультанты используют множество коротких вопросников и оценочных шкал (*rating scales*) для экспресс-оценки тех многообразных проблем, с которыми они сталкиваются в своей практике. Многие из этих средств оценки описаны в справочнике J. Fischer, & Corcoran (1994).

Данный раздел этой главы сосредоточен, главным образом, на рассмотрении областей, которые требуют применения разнообразных тестов и других инструментов при проведении обследования отдельных лиц, в противоположность изолированному использованию какого-то одного инструмента. Такого рода психологические оценки проводятся психологами с разной специализацией, хотя большинство из них получило ту или иную подготовку в области медицинской психологии и психологии консультирования. Клинические психологи традиционно проводили оценки в условиях психиатрических учреждений с целью постановки диагноза, составления прогноза или принятия терапевтических решений (Butcher, 1995; Hersen, Kazdin, & Bellack, 1991; Hurt, Reznikoff, & Clarkin, 1991; Maruish, 1994). Что касается консультирующих психологов, то первоначально они специализировались исключительно в области оценки профпригодности и профориентации (S. D. Brown, & Lent, 1992; Drummond, 1996; Gelso, & Fretz, 1992). Другие важные области, где практикуется психологическая оценка — такие, как школьная и судебная психология,[1] — получают свое определение исходя из среды, в которой работают профессиональные психологи (Shapiro, 1991; Vance, 1993; Weiner, & Hess, 1987). Дополнительные сферы деятельности, в которых применяются методики психологической оценки, определяются по типам исследуемых проблем, — например, психология здоровья (N. Adler, & Matthews, 1994; Streiner, & Norman, 1995; S. E. Taylor, 1990) и нейропсихология, занимающаяся изучением связей между мозгом и поведением (Benton, 1994; Maruish, & Moses, 1997). Области применения психологической оценки могут выделяться и по типу обслуживаемых пациентов или клиентов, — например, дети, супружеские пары или семьи (Conoley, & Werth, 1995; Fruzzetti, & Jacobson, 1992; Kamphaus, & Frick, 1996). Многие практикующие психологи работают в еще более специализированных областях, таких как педиатрическая нейропсихология (Batchelor, & Dean, 1996) или медицинская реабилитация (Cushman, & Scherer, 1995). Все эти области и специализации непрерывно эволюционируют. Вдобавок ко всему, границы между некоторыми сферами деятельности и специальностями стали гораздо менее четкими, а в ряде случаев, практически неразличимыми.

Замечательный пример такого частичного перекрытия сфер деятельности — растущее сближение между клинической психологией и психологическим консультирова-

[1] Судебная психология — это область практического применения психологии в судах, действующих по нормам статутного и общего права, причем психологи могут участвовать как в решении правовых вопросов, так и работать непосредственно с лицами, в отношении которых вынесено судебное решение или приговор о лишении свободы.

нием (см., например, Anastasi, 1979, 1990a; May, 1990).[1] Все чаще можно встретить психологов-консультантов, которые занимаются частной практикой, работают в клиниках и больницах. В то же время некоторые клинические психологи сейчас находят применение своим навыкам в оценке карьеры и отборе персонала (Lowman, 1989, 1991, 1993). Зримым признаком происходящих в настоящее время перемен служит то обстоятельство, что в проекте новых *Стандартов тестирования* вопросы использования тестов в медицинских, консультационных и учебных заведениях помещены в одну главу под названием «Психологическое тестирование и оценивание», тогда как в *Стандартах* 1985 г. клиническое тестирование и использование тестов в консультировании рассматривается по отдельности (см. главу 1).

Психологическая оценка. Что, в таком случае, позволяет говорить о психологической оценке как едином понятии, коль скоро она применяется разными специалистами в разных обстоятельствах при решении разных проблем у разных популяций? Одним из ее основных отличительных признаков является сфокусированность на интенсивном изучении одного человека или нескольких лиц (например, супружеской пары или семьи) по данным, полученным из многих источников. Посредством установления и поддержания раппорта клиницист может получить от клиента важные сведения о его прежней жизни, которые затруднительно добыть иными способами. Такие биографические данные обеспечивают надежную основу для понимания индивидуума и предсказания его последующего поведения. Кроме того, клиницист способствует процессу установления фактов тем, что играет роль стимула в ситуации межличностного взаимодействия. В этом смысле клиническое интервью выполняет функции ситуационного теста или имитационного моделирования, обеспечивая выборку межличностного поведения клиента в более или менее контролируемых условиях.

Информация, извлекаемая из данных наблюдения, интервью и истории болезни, объединяется с показателями тестов для получения целостной картины индивидуума (Beutler, & Berren, 1995). Клиницисты, таким образом, гарантируют себя от неправомерных обобщений на основе тестовых показателей. Вероятно, этим и объясняется, по крайней мере частично, продолжающееся использование ряда слабых или необоснованных, с психометрической точки зрения, тестов. До тех пор пока такие тесты служат источником гипотез для опытного клинициста, проверяемых им на других данных, сохранение этих инструментов может быть оправдано. Разумеется, существует опасность, что недостаточно опытный и излишне увлеченный тестированием клиницист, забывая об ограничениях используемого инструмента, может придавать неоправданно большое значение тестовым показателям.

Другой определяющий признак психологической оценки заключен в ее цели, которая обычно состоит в том, чтобы помочь в принятии обоснованных решений, касающихся дифференциального диагноза, выбора профессии, терапевтических рекомендаций, планирования образования, установления опеки над ребенком, виновности и

[1] В связи с этим обстоятельством и по соображениям удобства термин «клиницист» (*clinician*) будет употребляться в данном разделе для обозначения любого специалиста, связанного с проведением психологической оценки. Тем не менее полезно иметь в виду, что эти специалисты могут иметь базовое образование в какой-либо одной из указанных областей, и потому их ориентации и область опыта существенно различаются. Что касается сведений об использовании тестов клиническими, консультирующими, промышленными (организационными) и школьными психологами в их традиционных местах работы, см. P. S. Wise (1989).

многих других практически важных вопросов, затрагивающих одного или нескольких человек. Принятие решений происходит как естественное завершение процесса сбора, анализа, объединения различных данных и вдумчивого сообщения клиенту релевантной информации, касающейся его поведения. Сердцевину процесса психологической оценки составляет непрерывный цикл построения и проверки гипотез в отношении конкретного пациента или клиента. Каждая порция информации — будь это факт из истории болезни, мнение клиента или тестовый показатель — дает возможность выдвинуть гипотезу в отношении индивидуума, которая получит подтверждение или будет опровергнута по мере сбора других данных. В связи с этим не следует забывать, что любой отдельный источник данных, сколь бы надежным он ни казался, временами может давать неточную информацию.[1]

По существу, все действия, связанные с психологической оценкой, — от ясного объяснения ее основной цели до сообщения результатов — заключают в себе профессиональное *суждение (judgment)*, основывающееся на актуальном знании специфических проблем и категорий нуждающихся в помощи людей. Кроме того, применение индивидуальных тестов и других средств оценки требует специальных навыков их использования и должного внимания к их характеристикам, рассматриваемым в аспекте цели и контекста проводимой оценки (Drummond, 1996; C. T. Fischer, 1985; Groth-Marnat, 1990; G. Goldstein, & Hersen, 1990; Hood & Johnson, 1997; Maloney & Ward, 1976; Tallent, 1992; Walsh, & Betz, 1995). В этом отношении клинический метод всегда противопоставлялся проведению объективных, стандартизованных тестов и использованию статистических или актуарных методов для объединения данных посредством уравнений регрессии, множественных (составных) критических показателей и других «механических формул» (глава 6; см. также Dawes, Faust, & Meehl, 1993; L. R. Goldberg, 1991; Kleinmuntz, 1990; Wedding, & Faust, 1989).

Экологический подход, настаивающий на необходимости учета условий и обстоятельств жизни конкретного человека, оказал столь же значительное влияние на развитие психологической оценки, как и в случае возрастной психологии вместе с родственными ей областями (см., например, Moen, Elder, & Lüscher, 1995). Аналогично этому, возросшее понимание роли культуры во всяком поведении — включая проблемы, заставляющие людей искать помощи у специалистов по психическому здоровью, — стимулировало заинтересованность специалистов в получении информации о руководящих принципах проведения культурно-ориентированной оценки (см., например, American Psychiatric Association, 1994, p. xxiv and 843–849; Dana, 1993, 1996; Prediger, 1993; Suzuki et al., 1996).

Тесты интеллекта в контексте индивидуальной оценки. Такие тесты, как шкалы Векслера и Стэнфорд—Бине (см. главу 8), являются, по существу, индивидуальными, клиническими инструментами. Когда внимательный и опытный клиницист активно контактирует с обследуемым человеком где-то в течение часа, требуемого на проведение теста, он способен узнать об этом человеке явно больше того, что сообщает о нем *IQ* или какой-либо другой единичный показатель. Это верно даже в тех случаях, когда

---

[1] Представительную выборку современных методологических проблем исследования психологической оценки можно найти в специальном выпуске журнала *Psychological Assessment* (September 1995, Vol. 7, # 3).

тест проводится техническим персоналом, при условии, что сохраняется полный протокол ответов тестируемого.

Помимо использования тестов интеллекта для оценки общего уровня интеллектуальной деятельности индивидуума клиницисты обычно исследуют паттерн, или профиль показателей теста в целях обнаружения значимых превышений и снижений интеллектуальных функций относительно статистической нормы. Анализ профиля снабжает данными, которые могут помочь в диагностике локальных поражений мозга и разнообразных форм психопатологии, по-разному влияющих на функционирование интеллекта. Шкалы Векслера оказались идеально подходящими для такого анализа профилей, поскольку оценки по всем субтестам этих шкал выражаются в непосредственно сравнимых стандартных показателях. С самого начала Векслер описал ряд способов диагностического использования своих шкал. За прошедшее с тех пор время некоторые клиницисты предложили дополнительные способы использования шкал Векслера, а анализ профилей был применен в работе с другими инструментами (Delaney, & Hopkins, 1987; Elliott, 1990b; Kaufman, 1990, 1994; Matarazzo, 1972; Sattler, 1988, 1992). В большинстве способов анализа профиля используются разновидности трех основных методов. Первый заключается в оценке величины *разброса (scatter)*, или степени вариации тестовых показателей индивидуума, включая различия Вербального и Невербального *IQ*, общий разброс субтестов, а также сравнения нормированных показателей отдельных субтестов со средними показателями различных групп субтестов, например вербальных или скоростных. Второй метод состоит в анализе выраженных особенностей профиля индивидуума в свете *базовых данных (base rate data)* о частоте встречаемости таких особенностей в нормативной группе. Третий подход основан на *паттернах показателей (score patterns)*, связанных с определенными клиническими синдромами, такими как болезнь Альцгеймера, конкретные виды неспособности к научению (*learning disabilities*) или тревожные состояния. Д. Векслер и другие исследователи описали паттерны высоких и низких субтестовых показателей, а также комбинации субтестов, характеризующие эти и другие расстройства (см., например, Kaufman, 1990; Matarazzo, 1972).

Проводившиеся на протяжении нескольких десятилетий исследования различных форм анализа паттернов показателей по шкалам Векслера так и не смогли обеспечить сколько-нибудь серьезной статистической поддержки их диагностической ценности.[1] Фактически, критики этого подхода атаковали почти каждый его аспект по тому или иному поводу (F. C. Goldstein, & Levin, 1985; Kavale, & Forness, 1984; Macmann, & Barnett, 1994a, 1994b; McDermott, Fantuzzo, & Glutting, 1990). Тем не менее, судя по неизменной популярности шкал Векслера в клинической практике и по обширной литературе, посвященной систематизации, продвижению и усовершенствованию использования анализа паттернов показателей, этот подход по-прежнему остается предпочитаемым в области интерпретации данных, получаемых с помощью тестов интеллекта.

На чисто качественном уровне любая неравномерность выполнения теста также может подсказать пути для дальнейшего анализа. Существенные подсказки клиницисту могут давать как форма, так и содержание ответов тестируемого. Эксцентричность, излишние уточнения (*overelaboration*) или чрезмерные упоминания о себе (*self-reference*), например, могут указывать на расстройства личности. Качественный анализ ошибочных и правильных ответов может дать полезные подсказки для понимания

---

[1] См. Anastasi (1985a) по поводу обсуждения некоторых методологических требований, которые необходимо учитывать при оценивании этих исследований.

подходов к решению задач (*problem-solving*), уровня концептуальных представлений или когнитивных стилей. Нетипичное содержание ответов на задания теста является дополнительным источником эвристической информации. Еще одним источником качественных данных, доступным клиницисту во время проведения индивидуального теста интеллекта, служит общее поведение тестируемого в ситуации тестирования. Примеры включают двигательную активность, речь, эмоциональные реакции и аттитюды тестируемых в отношении тестирующего, а также их обращение с тестовыми материалами и реагирование на обстановку тестирования. Как правило, в силу своего сугубо индивидуального характера, такие качественные признаки невозможно обосновать количественными методами, принятыми для измерений групповых тенденций. Тем не менее признание решающей роли, которую могут играть наблюдения за поведением тестируемых, вызвало появление ряда инструментов, предназначенных для систематизации, подсчета и интерпретации некоторых поведенческих реакций во время сеанса тестирования (см., например, Руководство по оценке поведения во время сеанса тестирования для *WISC-III* и *WIAT* [Glutting, & Oakland, 1992]).

Наглядный пример тонкой модификации клинического использования тестов интеллекта, объединяющей психометрические данные с качественными наблюдениями, дают работы Алана Кауфмана (Alan S. Kaufman). В своих книгах по «интеллектуальному» тестированию интеллекта (Kaufman, 1979, 1990, 1994) Кауфман демонстрирует во всех подробностях, как клиницист может интегрировать статистическую информацию о тестовых показателях со знанием закономерностей развития человека, теории личности и других областей психологического исследования. Кауфман подчеркивает важность учета навыков и внешних (посторонних) условий, могущих влиять на выполнение субтестов, а также необходимость в дополнительной информации, извлекаемой из других тестов, истории болезни и клинического наблюдения за поведением во время тестирования, на фоне которой только и должна осуществляться интерпретация паттернов показателей. Тестовые показатели, вместе с информацией из других источников, приводят к формулированию гипотез об обследуемом человеке, которые могут проверяться по мере накопления информации в ходе построения целостной картины индивидуального случая. Самой важной отличительной особенностью метода Кауфмана является требование индивидуализированных интерпретаций выполнения теста вместо унифицированного применения какого-то одного из существующих разновидностей анализа паттерна. Один и тот же паттерн показателей может приводить к совершенно разным интерпретациям для разных людей.

Описанный Кауфманом основной подход, несомненно, представляет собой важный вклад в клиническое использование тестов интеллекта (см., например, Roecker, 1995)[1]. Даже его критики признают, что этому подходу стали отдавать предпочтение при обучении тестированию интеллекта и что он стимулировал разработку значительной части имеющегося сейчас вспомогательного программного обеспечения для

[1] Что касается критики метода Кауфмана и его ответов на нее, см. Kaufman (1994, chap. 1). Одна проблема, связанная с рядом отрицательных рецензий на предлагаемый Кауфманом подход, состоит, по-видимому, в недостаточно обоснованном предположении критиков, что клиницисты будут использовать его для принятия решений исключительно на основе величины показателей и различий между ними. И хотя совершенно верно, что механическое применение способов анализа профиля чревато самыми серьезными заблуждениями, такое предположение критиков полностью противоречит рекомендациям самого Кауфмана, равно как и принципам сложившейся практики проведения надежной оценки (Moreland et al., 1995).

интерпретации тестов интеллекта (McDermott, Fantuzzo, Glutting, Watkins, & Baggaley, 1992). Подготовленные Сэттлером руководства (Sattler, 1988, 1992) также служат превосходными иллюстрациями комбинированного — психометрического и клинического — использования индивидуальных тестов интеллекта. Следует, однако, признать, что реализация этих подходов по силам лишь квалифицированному клиницисту, хорошо осведомленному в целом ряде областей психологии и не испытывающему дефицита времени. К тому же доступность машинных интерпретирующих программ, которые в некоторых отношениях, бесспорно, облегчают применение этих методов, может легко склонить торопящегося или менее осведомленного специалиста к прямому заимствованию предлагаемых такими программами интерпретаций данных.

Тем временем разрабатываются новые процедуры использования профилей тестов интеллекта, причем некоторые из этих процедур также сочетают элементы психометрического и клинического подходов. Одна из интересных инноваций, проходящая сейчас стадию испытаний, — получила название метода «базового профиля» (*core profile*). Мак-Дермотт, Глаттин и их коллеги (Glutting, McDermott, Prifitera, & McGrath, 1994, 1995; McDermott, Glutting, Jones, & Noonan, 1989), так же как и Дондерс (1996), применили разные типы кластерного анализа к данным стандартизации шкал Векслера. Целью в данном случае было выведение базовых типов профиля, которые могли бы помочь в классификации результатов теста и в проверке гипотез об их клиническом значении. Дополняющая линия исследований использует многомерное шкалирование для выявления прототипических профилей интеллектуальной способности в популяции (Davison, Gasser, & Ding, 1996) и нацелена, в конечном счете, на количественную оценку степени соответствия между полученным в результате тестирования профилем индивидуума и прототипическими профилями, выявляемыми данной батареей тестов. Хотя эти исследования, бесспорно, преследуют заманчивую цель, на данный момент они носят поисковый характер и потому не представляют пока никакой пользы для клинической практики (см., например, Ryan, & Bohac, 1994).

Нейропсихологическая оценка

Методологические проблемы диагностики мозговых повреждений. Начало изучения поведенческих проявлений мозговых травм связано главным образом с работами Курта Гольдштейна и его коллег в начале 1920-х гг. (Goldstein, & Scheerer, 1941). Опираясь на обширные наблюдения за солдатами, получившими черепно-мозговые ранения во время Первой мировой войны, К. Гольдштейн дал классическое описание нарушений интеллектуальной деятельности вследствие мозговой травмы. Среди основных симптомов были снижение способности к абстрактному мышлению и тенденция реагировать на посторонние стимулы, могущие помешать нормальному восприятию.

Интерес к повреждениям мозга у детей возник в конце 1930-х — начале 1940-х гг. после исследований Альфреда Штраусса и его коллег (Strauss, & Lehtinen, 1947; H. Werner, & Strauss, 1941, 1943). Эти исследователи отобрали группу умственно отсталых детей, чьи истории болезни указывали на повреждения мозга вследствие травм или инфекционных заболеваний, случившихся до, во время и вскоре после рождения. Описание поведения этих детей представляло собой развитие и уточнение взрослого синдрома, описанного К. Гольдштейном. Полученная система отличительных признаков интеллектуальных и эмоциональных расстройств была широко признана в качестве типичной *именно для ребенка* с мозговыми повреждениями. Среди прочих в эту

систему включались специфические перцептивные и концептуальные расстройства, сочетавшиеся с относительно высокой вербальной способностью, а также гиперактивность, отвлекаемость и агрессивность. На протяжении многих лет, как в исследованиях поведения детей с повреждениями головного мозга, так и в клинической работе с такими детьми, господствовала одномерная концепция «жесткой связи поведения с конкретными участками мозга» (*organicity*). Этот подход привел к поискам диагностических тестов мозговых механизмов как таковых, и попыткам разработать коррекционные или обучающие программы, пригодные вообще для детей с поражениями мозга.

Начиная с 1950-х гг. психологи начали все яснее сознавать, что повреждение головного мозга может проявляться в широком многообразии поведенческих паттернов, — и именно это понимание ускорило развитие клинической нейропсихологии, нацеленной на применение знаний о связях поведения с мозгом в диагностике и реабилитации лиц с поражениями мозга. Нет и не должно быть ни одного симптома или их совокупности, свойственных всем случаям повреждения мозга. В действительности, поражение мозга может вызвать прямо противоположный паттерн поведения у двух разных людей. Подобные данные вполне согласуются с огромным разнообразием самой органической патологии, лежащей в основе поведенческих проявлений. Значительный прорыв в области анализа связей между мозгом и поведением был совершен благодаря исследованиям Ральфа Рейтана (Ralph Reitan) и его коллег в медицинском Центре университета штата Индиана (см. Matarazzo, 1972, chap. 13; Reitan, 1955, 1966). Эти исследования показали, что поражения левого полушария чаще связаны с более низким Вербальным *IQ* относительно Невербального *IQ* по шкалам Векслера (В < Н). Противоположный паттерн (В > Н) преобладает в группах с поражением правого полушария и в группах с диффузными мозговыми поражениями.

Нейропсихологические исследования продолжились в направлении классификации сложных взаимодействий других переменных с поведенческими эффектами мозговой патологии (см., например, Kolb, & Whishaw, 1990). Накоплено большое количество данных о том, что *возраст* влияет на поведенческие эффекты, к которым приводит поражение мозга. Эти поведенческие эффекты, по-видимому, также зависят от того, чему индивидуум успел научиться и какого уровня интеллектуального развития смог достичь до случившегося с ним несчастья. Исследования дошкольников, например, показали, что на этом возрастном уровне повреждения мозга, в общем, шире затрагивают интеллектуальные функции, чем в более старшем возрасте.

Как было установлено, *хроничность (chronicity)* также влияет на выполнение теста и взаимодействует с эффектами возраста. Имеющиеся данные позволяют предположить, что время, истекшее с момента повреждения мозга, может быть связано не только с прогрессирующими физиологическими изменениями, но и со степенью восстановления поведения благодаря научению или компенсаторному приспособлению. Наконец, следует отметить, что в некоторых случаях ослабление интеллекта может быть *косвенным результатом* мозгового поражения. На всем протяжении развития индивидуума органический фактор и фактор опыта взаимодействуют. Некоторые из поведенческих проблем, включаемых в классическую картину детей с поражениями головного мозга, могут быть, к примеру, лишь косвенным эффектом фрустраций и межличностных проблем, испытываемых ребенком с органически обусловленной интеллектуальной недостаточностью. Разовьются или нет подобные поведенческие проблемы, может, таким образом, зависеть от аттитюдов и степени понимания ребенка, обнаруживаемых у родителей, учителей и других значимых лиц в его окружении.

Теперь уже совершенно очевидно, что мозговые поражения охватывают широкий диапазон органических нарушений с не менее широким диапазоном поведенческих проявлений. Выполнение теста людьми с поражениями мозга, можно полагать, будет меняться в зависимости от причины, степени и места повреждения головного мозга; возраста, в котором произошло повреждение; возраста, в котором оценивается поведение индивидуума, а также степени и типа полученной им помощи специалистов. Поэтому ожидать однородности в поведении лиц с мозговой патологией было бы, по меньшей мере, неразумно.

С другой стороны, одно и то же интеллектуальное нарушение или поведенческое расстройство — и один и тот же диагностический признак в выполнении теста — могут иметь органическую, эмоциональную или смешанную этиологию. Наглядный пример — устойчивое беспамятство (*forgetfulness*). Разнообразные формы амнезии могут быть симптомом одного из многочисленных видов деменций, с известными органическими причинами, или же симптомом депрессивных расстройств эмоционального генеза; в добавление к этому, начало органического расстройства памяти часто сопровождается депрессией, представляя собой смешанную клиническую картину. Оценка нарушений памяти и, в особенности, дифференциальная диагностика деменции и депрессии у пожилых лиц относятся к числу наиболее часто возникающих в клинической нейропсихологии вопросов (Butters, Delis, & Lucas, 1995; Poon, 1986; Reeves, & Wedding, 1994; Storandt & VandenBos, 1994). Факторы опыта, которые могут быть не связаны с повреждением мозга в одних случаях или, наоборот, могут быть в той или иной степени связаны с ним в других случаях, обычно еще больше осложняют диагностическую картину. Поэтому интерпретация любого диагностического признака в выполнении теста требует дополнительной информации о рождении и истории жизни конкретного человека. Например, для измерения величины когнитивной недостаточности, так же как и для оценки степени восстановления интеллектуальных функций, крайне важно располагать информацией о преморбидном уровне способности индивидуума (см., например, Matarazzo, 1990). Уровень образования часто используется в качестве грубого показателя преморбидного функционирования интеллекта, но были разработаны и дополнительные методы оценки, основанной на биографических данных и результатах посттравматического тестирования (Vanderploeg, 1994b).

Вообще, практика нейропсихологической оценки — одна из самых трудных задач, требующая в действительности применения знаний о поведении личности, познавательной деятельности, неврологии и общей физиологии человека, причем не только в норме, но и в патологии. Поэтому неудивительно, что научная литература в этой области, так же как учебники и справочные руководства для студентов и практиков продолжают накапливаться с огромной скоростью. Один исчерпывающий компендиум сведений научного и клинического характера — многотомное *Руководство по нейропсихологии (Handbook of Neuropsychology)* под редакцией Боллера и Графмана (Boller, & Grafman, 1988–1995). Недавно был подготовлен целый ряд дополняющих это издание справочных работ по общей практике нейропсихологической оценки и программам вмешательства (Adams, Parsons, Culbertson, & Nixon, 1996; Golden, Zillmer, & Spiers, 1992; Lezak, 1995; Touyz, Byrne, & Gilandas, 1994; Vanderploeg, 1994a; R. F. White, 1992). Более специализированные труды по таким темам, как судебная нейропсихология (Valciukas, 1995), нейропсихологическая оценка профессионального нейротоксического поражения (Agnew, & Masten, 1994) и нейропсихологическое оценивание испаноязычных американцев (Ardila, Rosselli, & Puente, 1994), наряду со многими другими, продолжают издаваться быстрыми темпами.

Нейропсихологические инструменты. Довольно много тестов специально разрабатывалось в качестве клинических инструментов для оценки нейропсихологического повреждения[1] (см., например, Lezak, 1995; Spreen, & Strauss, 1991). Эти тесты часто рассматриваются как индикаторы органических (*organicity*) или мозговых поражений. Среди оцениваемых с помощью таких инструментов функций главными считаются наиболее чувствительные к патологическому процессу, а именно восприятие пространственных отношений и память на только что заученный материал. Зрительно-моторный гештальт тест Бендер (*Bender Visual Motor Gestalt Test*), обычно называемый Гештальт тестом Бендер (*Bender-Gestalt Test* [*BGT*] — Bender, 1938; Canter, 1996; Heaton, Baade, & Johnson, 1978; Koppitz, 1964, 1975; Pascal & Suttell, 1951), и Тест визуальной ретенции Бентона (*Benton Visual Retention Test–Fifth Edition* [*BVRT*] — Sivan, 1991) служат примерами инструментов этого типа и использовались в качестве скрининговых тестов на протяжении многих десятилетий. Однако вследствие широкого разнообразия органических мозговых дисфункций и сопутствующей им многообразной поведенческой дефицитарности ни один тест не может считаться достаточным для скрининга мозговых поражений вообще, а такие однопрофильные тесты, как *BGT* и *BVRT* еще менее пригодны для дифференциальной диагностики.

Клинические нейропсихологи часто используют комбинацию доступных тестов, оценивающих различные навыки и дефициты, в манере, получившей название метода «гибкой батареи» (см., например, Bauer, 1994; Goodglass, 1986). У этой процедуры есть преимущество в том, что она позволяет получать комбинации тестов, которые специально подбираются к наличным проблемам каждого отдельного пациента. Но у нее есть и ряд ограничений. Вероятно, при таком подходе может происходить излишнее дублирование функций некоторых включаемых в батарею тестов, тогда как некоторые критические области могут оказаться пропущенными. Предварительный отбор тестов, подходящих для конкретного больного, становится серьезным испытанием мастерства и проницательности клинициста. Кроме того, самостоятельно разработанные тесты чаще всего несопоставимы с нормами и показателями других шкал. Эмпирических данных о взаимосвязях различных тестов, по-видимому, также недостаточно. Все это существенно затрудняет интерпретацию результатов на основе паттернов показателей.

По этим причинам систематически предпринимались попытки составить комплексные стандартизованные батареи, которые обеспечивали бы измерение всех важных — с нейропсихологической точки зрения — умений и навыков. Такая батарея может выполнять несколько функций. Она позволяет обнаруживать поражения мозга с высокой степенью успешности, а также помогает идентифицировать и локализовать пораженные участки мозга. Кроме того, она дифференцирует специфические синдромы, связанные с церебральной патологией. Наконец, такая батарея помогает спланировать восстановительное обучение благодаря выявлению конкретных типов и степени поведенческой дефицитарности. Два основных примера комплексных нейропсихологических батарей — Батарея нейропсихологических тестов Халстеда—Рейтана (*Halstead-Reitan Neuropsychological Test Battery* [*HRB*] — Reitan, & Wolfson, 1993) и

[1] Тесты этой категории регулярно рецензируются в *Ежегодниках психических измерений (Mental Measurements Yearbooks)*, а их текущий перечень приведен в *TIP-IV*, p. 1116. (Полезную информацию об отечественных разработках в этой области читатели могут найти в книге: *Вассерман Л. И., Дорофеева С. А., Меерсон Я. А.* Методы нейропсихологической диагностики: Практическое руководство. — СПб.: Стройлеспечать, 1997. — *Примеч. науч. ред.*)

Нейропсихологическая батарея Лурия—Небраска (*Luria-Nebraska Neuropsychological Battery* [*LNNB*] — Golden, Purisch, & Hammeke, 1985). Обе эти батареи, в том числе и их детские версии, имеют одинаковое назначение, но при этом различаются в нескольких важных отношениях. *HRB*, изданная раньше *LNNB*, была разработана Рейтаном на основе исследований Халстеда (Halstead, 1947) и предоставляет клиницистам некоторую свободу в выборе количества и числа тестов при обследовании конкретного пациента (что касается рецензий, см. Dean, 1985; M. J. Meier, 1985).[1] При разработке LNNB был использован ряд теорий и диагностических методик А. Р. Лурия (Christensen, 1975; Luria, 1973, 1980). Эта батарея более стандартизована в том, что касается содержания, стимульных материалов, предъявления тестов и подсчета показателей, и, по сравнению с *HRB*, требует существенно меньше времени для проведения обследования (что касается рецензий, см. J. H. Snow, 1992; Van Gorp, 1992).

Современные достижения в области прямой оценки повреждений головного мозга с помощью электроэнцефалографии и таких методов нейроинтроскопии, как ЯМР-интроскопия и позитронная эмиссионная томография, оказывают глубокое влияние на общую и клиническую нейропсихологию. Хотя доступные технологии непрерывно совершенствуются, ни один из диагностических методов не дает 100 % надежности. В большинстве случаев нейропсихологи работают совместно с неврологами и другими специалистами, чтобы иметь возможность получать подтверждающую информацию из разных источников. В условиях клиники тщательно стандартизованные средства измерения поведения выполняют вместе с другими методами, важную функцию в оценке локальных поражений мозга, а также в планировании и мониторинге восстановительного обучения. В области фундаментальной науки интеграция нейропсихологических и нейроинтроскопических методик обещает обернуться небывалым прогрессом в понимании связей мозга и поведения (см., например, Gur, & Gur, 1991, 1994).

Выявление недостаточной специфической обучаемости (НСО)[2]

Семидесятые годы ознаменовались интенсивной разработкой программ диагностики и коррекции НСО. Педагоги стали все острее сознавать широкую распространенность этого препятствия среди школьников и даже среди студентов и других взрослых (см., например, Gregg, Hoy, & Gay, 1996; Kravets, & Wax, 1992; Wang, Reynolds, &

[1] Рассел и Старки (Russell, & Starkey, 1993) разработали комплексную автоматизированную систему для сбора первичных показателей по многим тестам *HRB* (и другим методикам) и их преобразования в показатели единой шкалы с учетом возраста и образования. Их программа, кроме того, строит профили и составляет краткие сводки результатов в целях облегчения интерпретации. Несколько раньше аналогичная система была подготовлена в печатном варианте (Heaton, Grant, & Matthews, 1991) и недавно также реализована в виде компьютерной программы.

[2] К сожалению, в отечественной психологии до сих пор не найден устойчивый точный термин для перевода английского выражения (*specific*) *learning disabilities*. Встречающиеся варианты перевода — от совершенно неопределенного «(специфические) трудности в обучении» до буквального «неспособность к об/на/учению» — можно рассматривать лишь как временные, требующие к тому же пояснений всякий раз, когда они употребляются в разных контекстах. В такой ситуации выражение *specific learning disabilities* можно переводить по-разному, в зависимости от контекста, что и делается в этом издании (с указанием в скобках термина на языке оригинала). Однако в разделе, специально посвященном этой теме, я все же решился предложить свой вариант перевода — «недостаточная специфическая обучаемость (НСО)», — руководствуясь желанием как можно точнее передать содержание в стилистически удобной форме. Разумеется, это лишь один из многих возможных вариантов, а насколько он удачен, покажет время. — *Примеч. науч. ред.*

Walberg, 1991), хотя большое число относимых к этой категории лиц может, отчасти, отражать ошибки классификации вследствие нестрогого употребления термина НСО. Варьирование применяемой в этих случаях терминологии отражает как меняющиеся со временем подходы к данной проблеме, так и различия между медицинской, педагогической и психологической ориентацией в отношении НСО (см., например, American Psychiatric Association, 1980, 1994). Согласно определению, сформулированному федеральным правительством в параграфах 94–142 Свода законов общественного права (Public Law 94–102), где речь идет об обеспечении образования для детей с умственными и физическими недостатками, *недостаточная специфическая обучаемость (specific learning disability)* описывается как

> ...расстройство одного или более основных психических процессов, ответственных за понимание или использование речи, устной или письменной, которое может проявиться в недостаточной способности слушать, думать, говорить, читать, писать, произносить по буквам или производить математические вычисления. Данный термин включает такие состояния, как недостатки восприятия (*perceptual handicaps*), повреждение мозга, минимальная мозговая дисфункция, дислексия и связанная с развитием афазия (*developmental aphasia*). Данный термин не относится к детям, чьи учебные проблемы имеют своей первичной причиной дефекты зрения, слуха или моторики, психическую задержку, эмоциональное расстройство и неблагоприятные средовые, культурные или экономические условия жизни (*Federal Register*, 1977, p. 65 083).

Далее в этих же параграфах (P. L. 94–142) уточняется, что диагноз НСО следует применять только к детям, которые 1) обнаруживают «резкое несоответствие» интеллектуальной способности достигнутому уровню навыков коммуникации и математических действий и 2) не могут овладеть ими на уровне, соответствующем их возрасту и интеллектуальной способности, даже при обеспечении должного обучения.[1] Как следует из этого определения, диагноз НСО не следует применять до тех пор, пока не исключен ряд других состояний и условий как возможных причин учебных или психологических трудностей конкретного ребенка. С годами приходило все большее понимание того, что неоднородная совокупность лиц с НСО состоит из подгрупп, которые поддаются дифференциации в виде кластеров симптомов (Rourke, 1990; Feagans, Short, & Meltzer, 1991; Geary, 1993; S. R. Hooper, & Willis, 1989; Pennington, 1991; Shankweiler et al., 1995; H. L. Swanson & Keogh, 1990). Тем не менее, и сейчас существуют существенные расхождения в теоретической ориентации даже между специалистами по НСО. Эти расхождения отражаются как в инструментах тестирования, так и в коррекционных программах.

Обычно дети с НСО демонстрируют нормальный интеллект, нередко даже превышающий средний уровень, в сочетании с выраженными трудностями в овладении одним или несколькими основными школьными навыками (наиболее часто — чтением). Следует, однако, заметить, что НСО может встречаться на любом интеллектуальном уровне, даже если дети с НСО, сочетающейся с задержкой психического развития, не подходят под юридическое определение НСО. Дети с НСО проявляют также

[1] Более подробное обсуждение положений этого закона (P. L. 94–142) и его следствий для диагностики НСО см. в работах C. R. Reynolds (1990, 1992b). Сэттлер (Sattler, 1988, p. 598–617) дает краткий, но весьма информативный обзор разновидностей НСО и методов их оценки.

различные поведенческие симптомокомплексы. Главными среди таких симптомокомплексов являются трудности в восприятии и кодировании информации, недостаточная интеграция входных сигналов разной модальности и нарушение сенсомоторной координации. Нарушение языкового развития типичны для детей с НСО. Ограниченность памяти, произвольного внимания и навыков отвлеченного мышления тоже часто встречается у таких детей, как и некоторые эмоциональные и мотивационные симптомы. В частности, агрессия, а также другие эмоциональные и межличностные проблемы вполне могут развиться как прямая реакция на неудачи ребенка в учении и фрустрации, вызванные его НСО. При оценке поведения ребенка надо иметь в виду, что многие специфические трудности, нормальные для раннего возраста (скажем, в 3 года), становятся признаками дисфункции, если сохраняются в старшем возрасте. Поэтому существует потребность в системе координат возрастного развития, хотя бы с качественными, если уж не с количественными нормативами.

Методики оценки. Независимо от теоретических позиций, все специалисты признают, что выявление НСО требует широкого ассортимента тестов, дополняемых процедурами наблюдения. Это вытекает, по меньшей мере, из трех характерных особенностей диагностической задачи: 1) разнообразия поведенческих расстройств, связанных с НСО; 2) индивидуальных различий в конкретном сочетании симптомов; 3) потребности в предельно конкретных сведениях о природе и степени НСО для каждого случая.

Обычно оценка детей с НСО обеспечивается совместными усилиями группы специалистов. Работающий с классом учитель может проводить групповые тесты и применять другие скрининговые или «широкодиапазонные» (*wide-band*) инструменты. Регулярно применяемые батареи достижений также можно использовать в этой связи, особенно предназначенные для младшего школьного и дошкольного уровней и допускающие критериально-ориентированный (т. е. содержательно-ориентированный) анализ конкретных сильных и слабых сторон. Для этой цели подходят некоторые из инструментов, упоминавшихся в разделе этой главы, посвященном тестированию в образовании.

Для оценки НСО особенно подходят проводимые индивидуально, «широкодиапазонные» тесты достижений. Эти тесты могут даваться учителем, хотя клиницист лучше справится с дополнительными качественными наблюдениями и интерпретацией показателей. Один из инструментов этой категории — Тест учебных достижений Кауфмана (*K-TEA* — Kaufman, & Kaufman, 1985), проведение которого с использованием распространяющегося комплекта типа пюпитра показано на рис. 17–4. Ряд батарей включают как субтесты, относимые к разряду мер частных способностей (*aptitudes*) или общей когнитивной способности (*cognitive ability*), так и субтесты, считающиеся мерами достижений. К числу наиболее полных тестов из этого класса относится Пересмотренная психо-педагогическая батарея Вудкока—Джонсона (*Woodcock-Johnson Psycho-Educational Battery–Revised* [*WJ-R*] — McGrew, 1994; McGrew, Werder, & Woodcock, 1991). Такие тесты, как Стэнфорд—Бине, K-ABC и шкалы Векслера, обеспечивают получение не только общего показателя типа *IQ*, который помогает отличить психическую задержку от НСО, но и богатой качественной информации, касающейся различных специфических «неспособностей» (Kaufman, 1990; 1994; Sattler, 1988, 1992). Например, эти тесты могут выявлять возможные недостатки восприятия и узнавания зрительно предъявляемых паттернов, двигательные затруднения, ограничения крат-

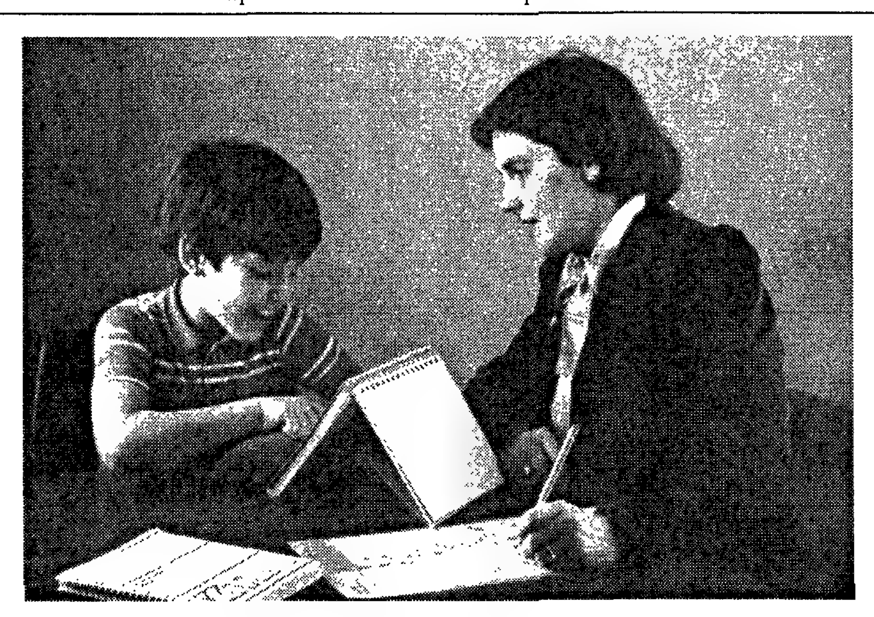


Рис. 17-4. Проведение Теста учебных достижений Кауфмана (*K-TEA*) с использованием комплекта типа пюпитра. Скрепленные кольцами стимульные карточки образуют свободно устанавливаемый пюпитр, позволяющий одновременно предъявлять каждое задание тестируемому и соответствующие указания тестирующему

(С любезного разрешения American Guidance Service)

ковременной памяти, неспособность оперировать абстрактными понятиями и многие виды речевых расстройств. Издатели шкал Векслера разработали также Индивидуальный тест достижений Векслера (*Wechsler Individual Achievement Test* [*WIAT*] — Psychological Corporation, 1992) — комплексную батарею, увязанную с векслеровскими шкалами интеллекта и предназначенную служить их дополнением при оценке НСО.

Хотя существующие тесты могут быть полезными в реализации современных принципов оценки НСО, многие исследователи неоднократно заявляли о потребности в новом, более информативном подходе к диагностике и оцениванию детей с НСО. Все они указывали на необходимость более ясного теоретического обоснования процедуры проведения оценки и более глубокого понимания процессов, специфичных для каждого случая, для повышения эффективности оценки и соответствующих программ коррекции (см., например, Das, Naglieri, & Kirby, 1994; R. B. Kline, Snyder, & Castellanos, 1996; C. R. Reynolds, 1992b).

Динамическая оценка. Термин «динамическая оценка» (*dynamic assessment*) охватывает множество разнообразных клинических методик, которые предполагают по существу намеренное отступление от стандартизованного или единого для всех теста для получения дополнительных качественных данных об индивидууме. Несмотря на то что квалифицированные клиницисты пользовались такими методиками и раньше, популярность этого подхода начала расти начиная с 1970-х гг. (Lidz, 1981, 1987, 1991, 1995). Он служил способом получения дополнительных данных, причем не только в случаях НСО, но и при работе с другими детьми, испытывавшими трудности в обучении, например, вследствие слабой или умеренной психической задержки. Была также

проверена, в предварительном порядке, полезность этого подхода для оценки одаренности, особенно у детей, растущих в экономически неблагоприятных условиях (см., например, Bolig, & Day, 1993).

Одна из первых таких качественных адаптаций процедуры тестирования получила название «тестирование пределов» (*testing the limits*). При этой процедуре тестируемому могут даваться дополнительные сведения или, иначе говоря, подсказки, — и чем больше подсказок требуется для удовлетворительного выполнения заданий, тем больше выражена НСО. Изменения стандартной процедуры тестирования, допускаемые при «тестировании пределов», сходны с теми, которые имеют место в некоторых специальных адаптациях, применяемых в тестировании лиц с физическими недостатками, и предполагают те же ограничения и предосторожности в отношении интерпретации итогового выполнения теста (см. главу 9).

Несколько позднее был разработан подход, названный *оценкой потенциала обучения (learning-potential assessment* — Babad, & Budoff, 1974; Campione, & Brown, 1979; Feuerstein, 1979; Glutting, & McDermott, 1990; Hamers, Sijtsma, & Ruijssenaars, 1993; Lidz, 1991). Термин «потенциал» в этом названии может быть отнесен к необоснованному предположению, будто исследуемая способность существовала всегда и ее нужно только «раскрыть». Однако на самом деле эти процедуры имеют структуру типа «тест—обучение—тест» и заключаются в обучении учащегося различным средствам выполнения задания, с которым он первоначально не смог справиться. Внешне эта процедура напоминает способ, используемый в некоторых прогностических образовательных тестах, где испытуемым дают выборочную задачу, требующую такого рода научения, с которым им предстоит столкнуться в конкретном учебном курсе. Тем не менее процедура оценки потенциала обучения отличается от методики проведения прогностических тестов, по крайней мере, в двух отношениях: 1) обследуемому ученику дают указания или индивидуальные советы, а 2) используемые задания обычно требуют более широких учебных умений или навыков решения задач (*problem-solving*).

Методики динамической оценки, начало которым было положено работами Фейерстейна и др., открывают ряд перспектив. Связывая оценку и обучение, они стимулируют исследования пределов изменяемости академической способности и содействуют разработке программ оптимальной коррекции. В добавление к этому, они дают в руки квалифицированного клинициста средство оценки, позволяющее получать более ясные описания когнитивной деятельности и ее чувствительности к корригирующим вмешательствам, чем стандартизованные тесты интеллекта. Несмотря на эти преимущества, динамическим методикам не удалось избежать критики. Прежде всего, сомнению подвергается их *переносимость (transportability)* или, иначе говоря, то, в какой степени их могут эффективно использовать разные клиницисты. Другое сомнение касается *распространимости (generalizability)* эффектов коррекции, полученных с этими весьма широкими заданиями (обычно с задачами на невербальное рассуждение типа используемых в Прогрессивных матрицах Равена или в Невербальной шкале Векслера), на выполнение реальных учебных заданий в школе и дома. Кроме того, хотя большинство сторонников динамической оценки были бы не прочь избавиться от тестовых показателей, которые являются типичными «статическими» когнитивными мерами (как, например, показатели тестов интеллекта), задача документирования изменений без использования чисел оказалась для них трудноразрешимой (A. L. Brown et al., 1992; R. E. Snow, 1990). Тем не менее поисковые исследования потенциальной полезности методик динамической оценки продолжают идти быстрыми тем-

пами. Несколько недавно опубликованных работ, открывающих в этом отношении некоторые перспективы, посвящено прояснению роли различных учебных стратегий в усвоении математических понятий и разработке автоматизированной системы динамической оценки в области решения задач на умножение многоразрядных чисел (Gerber, Semmel, & Semmel, 1994; Jitendra, Kameenui, & Carnine, 1994).

С другой стороны, разработанная Эмбретсоном (Embretson, 1987, 1990, 1992) многомерная модель изменения латентных черт обходит многие — практически непреодолимые — технические трудности измерений изменения выполнения заданий или, по-другому, научения, свойственные традиционным тестам (см., например, Cronbach, & Furby, 1970). Подход Эмбретсона опирается на теорию «задание—ответ» (или *IRT*, рассмотренную в главе 7) и компьютеризованное адаптивное тестирование (или *CAT*, обсуждавшееся в главе 10), чтобы обойти эти проблемы путем тестирования каждого индивидуума с помощью заданий, близких к его пороговому уровню, где эффекты научения максимальны. Задания этого уровня дают еще и наиболее надежную оценку деятельности каждого тестируемого. Кроме того, за счет применения методов декомпозиции задачи, разработанных в когнитивной психологии, Эмбретсон может систематически варьировать специфические когнитивные процессы, необходимые для решения задачи, которую ставит перед тестируемым каждое задание теста. Это может осуществляться, например, путем использования подзадач, выполнение которых требует лишь нескольких шагов полного процесса решения, или путем обеспечения обучения, влияющего на определенные аспекты выполнения задачи. Очевидно, что продолжающиеся исследования Эмбретсона вносят существенный вклад в психометрию, сводя воедино достижения в таких разных областях, как клиническая психология, экспериментальная когнитивная психология, математическая статистика и машинная технология (Embretson, 1993, 1995).

Поведенческая оценка

Разнообразные методики, относимые к общему понятию модификации поведения, отображают прямое использование основных закономерностей научения в практическом управлении изменением поведения. В основном эти методики заключаются в применении принципов обусловливания для приобретения или усиления желательного и устранения нежелательного поведения. Терапия поведения (*behavior therapy*) постепенно расширялась, чтобы охватить увеличивающееся многообразие психологических проблем, непрерывно пополняющийся репертуар методик вмешательства и прибавление к внешним моторным реакциям когнитивных и эмоциональных реакций (Bandura, 1969, 1986; Hersen et al., 1991; Lazarus, 1981).

Методики оценки. В ранних приложениях терапии поведения практически не уделялось внимания оценке. Однако с середины 1970-х гг. важность процедур оценки начала все больше и больше признаваться специалистами (Barrios, 1988; A. S. Bellack, & Hersen, 1988; Haynes, 1991; Mash, & Terdal, 1988; Nelson, & Hayes, 1986; O'Brien, & Haynes, 1993; Ollendick, & Hersen, 1993).[1] Главные функции, выполняемые методика-

---

[1] Инструменты этой категории регулярно рецензируются в *Ежегодниках психических измерений (Mental Measurement Yearbooks)*, а их текущий перечень приведен в *TIP-IV*, р. 1095–1096). Что касается информации об использовании машинной технологии в разных типах методик поведенческой оценки, см. Kratchowill et al. (1991).

ми оценки в терапии поведения, можно разнести по трем рубрикам. Во-первых, методики оценки помогают *определить проблему конкретного человека (defining the individual's problem)* посредством функционального анализа релевантного поведения. По существу, данный анализ подразумевает полную спецификацию цели лечения, такой как преодоление фобии или навязчивых мыслей, и включает описание стимулов, вызывающих целевое поведение, ситуаций, в которых оно возникает, а также характера, величины и частоты специфических реакций. Во-вторых, методики оценки могут давать терапевту ориентиры в том, что касается *подбора подходящих способов терапии (selecting appropriate treatments)*. В-третьих, всегда есть потребность в *оценивании изменений поведения (assessing the behavior change)*, происходящих в результате терапии. Такие поведенческие оценки должны включать как методики для мониторинга изменений, чтобы можно было оценить эффективность терапии и ввести, если это необходимо, процедурные изменения, так и средства итоговой оценки (*terminal measures*), позволяющие установить достижение удовлетворительного состояния и спланировать, при необходимости, дополнительные процедуры.

При рассмотрении конкретных методик оценки следует учитывать, что, во-первых, одни и те же методики часто могут давать информацию, релевантную всем трем функциям. Во-вторых, выбор методик оценки зависит от существа проблемы, особенностей клиента (включая характеристики среды, в которой он должен действовать) и оборудования, имеющегося в конкретной клинике. В-третьих, в подавляющем большинстве случаев желательно использовать комбинацию нескольких методик оценки.

В свою очередь, доступные пользователям методики поведенческой оценки можно также разделить на три типа: самоотчеты клиента, непосредственное наблюдение поведения и физиологические замеры. Ходя далеко не каждый центр может предоставить оборудование, необходимое для проведения *физиологических замеров (physiological measures)*, именно такие замеры дают дополнительные объективные данные при оценивании некоторых состояний, таких как тревога, расстройства полового возбуждения и расстройства сна (Sturgis, & Gramling, 1988). Примеры включают замеры электродермальной (кожно-электрической), мышечной и электроокуломоторной активности, а также функционирования сердечно-сосудистой системы, половых органов и головного мозга.

Непосредственное наблюдение (direct observation) может проводится в естественной обстановке родителями, учителями, персоналом лечебных, исправительных или других учреждений и специально подготовленными наблюдателями. При этом могут использоваться такие вспомогательные средства, как контрольные таблицы (или схемы наблюдения), оценочные шкалы и дневные (суточные) графики. Таким наблюдениям присущ ряд недостатков (Barrios, 1993; см. также главу 16). По этой причине в клиниках часто используют моделирование значимых ситуаций. Одновременно имеет место тенденция к увеличению использования автоматических датчиков и контрольных устройств, которые обеспечивают непрерывную, объективную регистрацию поведения в реальных и вымышленных ситуациях (W. W. Tryon, 1985, 1991).

Категорию *самоотчетов клиента (self-report by the client)* составляют весьма разнородные методики. В их число входят клинические интервью, проводимые терапевтом, отчеты клиента о целевом поведении и связанных с ним состояниях на основе текущего самоконтроля, а также многообразные письменные отчеты по контрольным спискам и опросникам. Используются также некоторые опросники типа стандартизованных самоотчетов — в оригинальных или адаптированных версиях — как для пред-

варительного отсеивания и выявления, так и для отслеживания целевого поведения. К группе самых простых и наиболее широко используемых стандартизованных самоотчетов относится пересмотренный Опросник депрессии Бека *(Beck Depression Inventory* [*BDI*] — Beck, & Steer, 1993) — состоящая из 21 пункта шкала, предназначенная для оценки тяжести депрессии на основе самооценок.[1] Одним из самых новых стандартизованных инструментов является Инвентарь употребления алкоголя (*Alcohol Use Inventory* — Horn, Wanberg, & Foster, 1990) — опросник, состоящий из 228 пунктов и оценивающий то, насколько человек втянулся в употребление спиртного, исходя из довольно сложной концептуальной и психометрической модели с использованием параллельных шкал на разных уровнях обобщенности (что касается рецензий, см. Drummond, 1995; McNeely, 1995). Для использования в конкретных исследовательских проектах и терапевтических программах было подготовлено множество других инструментов, и, хотя некоторые из них еще не изданы для коммерческого распространения, они обычно полностью приведены и описаны в журнальных статьях или книгах (см., например, A. S. Bellack, & Hersen, 1988).

Совсем недавно были разработаны инструменты иного рода, которые содержат объективные рейтинговые шкалы, заполняемые несколькими информантами. Примером одного из них может служить упоминавшаяся в главе 16 Система для оценки поведения детей (*BASC*), разработанная Сесил Рейнолдс (Cecil Reynolds) и Рэнди Кауфманом (Randy Kauphman). Эта система включает оценочные шкалы для родителей и учителей, а также форму для кодирования и декодирования прямых наблюдений за поведением в классе; кроме того, в *BASC* входит форма самоотчета для детей и схема структурированного интервью для получения от родителей сведений о развитии ребенка. Рейтинговая система социальных навыков (*Social Skills Rating System* [*SSRS*] — Gresham, & Elliott, 1990) также предусматривает формы для родителей, учителей и учеников, на основе которых оценивается положительное и проблемное поведения учеников в школе и дома. Одной из особенно полезных особенностей *SSRS*, по мнению рецензентов, является наличие в этой системе компонента, позволяющего увязывать результаты оценки с планированием стратегий вмешательства (что касается рецензий, см. Benes, 1995; Furlong, & Karno, 1995).

Оценка карьеры

Практическая деятельность в сфере оценки карьеры (*career assessment*) состоит в том, чтобы помочь конкретному человеку сделать выбор наиболее подходящей профессии с учетом его способностей, интересов, целей, ценностей и темперамента, а также требований самой профессии. Очень мало сфер жизни человека может сравниться по важности с профессиональными занятиями людей, и не только потому, что большинство проводит за работой огромную часть отпущенного им времени, но и потому, что работа обычно создает возможности для множества внутренних и внешних вознаграждений (Super, & Šverko, 1995). В добавление к этому, происходящие в наше время быстрые перемены в характере и условиях работы заставляют гораздо большее число людей обращаться к выбору профессии не один, а несколько раз за свою жизнь. Поэтому неудивительно, что развитие теорий в области выбора и построения карьеры

[1] Обширная библиография по *BDI* и рецензии более ранней редакции этой шкалы помещены в 11-м выпуске Ежегодника психических измерений (*MMY*).

идет довольно быстрыми темпами. Уже в 1990-х гг. число принципиально новых теоретических представлений в данной области достигло или даже превысило их совокупное количество, накопившееся с 1950-х гг., когда Дональд Супер (Super, 1953) и Джон Холланд (Holland, 1959) предложили свои первые теоретические разработки проблемы выбора профессии и профессионального цикла.[1]

С точки зрения тестирования мы уже рассматривали те инструменты, которые находят самое прямое применение в профконсультировании, а именно инвентари интересов и комплексные батареи способностей, обсуждавшиеся в главах 14, 10 и в разделе данной главы, посвященном тестированию в сфере профессиональной деятельности. Выбор профессии часто предполагает выбор образа жизни, связанного с характерным набором ценностей. Поскольку инвентари интересов, по существу, оценивают систему ценностей индивидуума, их начинают все больше и больше рассматривать в качестве главных инструментов эффективного планирования карьеры. В этом разделе мы рассмотрим два более специализированных вида инструментов, которые специально разрабатывались для профконсультирования: комплексные программы изучения возможной карьеры (*comprehensive programs for career exploration*) и средства оценки готовности к трудовой деятельности (*measures of career maturity*). Характеристику и критический анализ гораздо большего числа инструментов всех типов для оценки карьеры можно найти в обязательном для каждого специалиста руководстве (Kapes, Mastie, & Whitfield, 1994), выпущенного уже третьим изданием.

Комплексные программы изучения возможной карьеры. Несколько комплексных батарей способностей было включено в состав систем профориентации. Примером могут служить описанные в главе 10 Дифференциальные тесты способностей (*DAT*), которые могут использоваться в сочетании с Инвентарем профессиональных интересов (*CII* — Psychological Corporation, 1991a, 1991b). Эти два инструмента разрабатывались совместно в целях облегчения сравнений результатов их применения в процессе профориентации.

Другой пример — программа, разработанная Управлением размещения и регулирования рабочей силы США (*USES*), чья Батарея тестов общих способностей (*GATB*) обсуждалась в разделе этой главы, посвященном профессиональному тестированию. К числу наиболее полезных инструментальных средств, созданных в рамках профориентационной программы *USES*, относятся Полное руководство по изучению возможной карьеры (*Complete Guide for Occupational Exploration* [*CGOE* — Farr, 1992]) и Усовершенствованное руководство по изучению возможной карьеры (*Enhanced Guide for Occupational Exploration* [*EGOE*] — Maze & Mayall, 1995). Предназначенные для использования консультантами, а также учащимися и ищущими работу людьми, эти руководства распределяют тысячи специальностей в мире труда по группам в соответствии с основными областями интересов, структурами способностей и другими требованиями для успешной профессиональной деятельности. Каждый человек может воспользоваться этими руководствами для предварительного изучения возможной

---

[1] Полезный обзор этих и других хорошо обоснованных теорий профессионального выбора, так же как и некоторых появляющихся теоретических разработок данной проблемы в более современном ключе, см. в книге Brown, Brooks, et al. (1996). Книга *Convergence in Career Development Theories* (Savickas, & Lent (Eds.), 1994) — другой ценный источник информации о теориях выбора и построения карьеры, в частности чем они похожи, чем различаются и в чем дополняют друг друга. Многие главы двух этих книг написаны авторами рассматриваемых в них теорий.

карьеры, выделяя те группы специальностей, к которым у него есть сильный интерес, а затем проверяя подготовку и навыки, которых они требуют. *CGOE* включает список из 12 741 специальности, полностью соответствующий перечню специальностей в Словаре названий профессий (*Dictionary of Occupational Titles* — U. S. Department of Labor, 1991), тогда как *EGOE* включает только 2800 специальностей, охватывающие 95 % рабочей силы, но дает больше информации по каждой из них.

Более современный подход к профконсультированию предусматривает процедуру объединения доступной информации из многих источников в единую комплексную программу изучения карьеры. Такая информация может включать показатели разнообразных тестов (каждый из которых имеет свои собственные нормативные и интерпретирующие данные), биографические сведения (включая уровень образования и опыт работы), а также данные о выраженных интересах, предпочтениях и системе ценностей индивидуума. Этот подход представлен, хотя и в различной степени, несколькими доступными пользователям инструментами, такими как Пересмотренная система принятия карьерных решений Харрингтона—О'Ши (Harrington, & O'Shea, 1993) и *ACT*-программа планирования карьеры (*American College Testing Career Planning Program* [*CPP*] – ACT, 1994).

Замечательным примером таких интегративных программ изучения возможной карьеры является пересмотренная версия Профориентационной диалоговой системы (*System for Interactive Guidance Information* [*SIGI-PLUS*]), упоминавшейся в главе 3. Используя интерактивную программу, *SIGI-PLUS* дает клиенту возможность проводить двустороннюю связь с компьютером — спрашивать и отвечать на вопросы, запрашивать и получать информацию. Программа связана с обширной базой данных о характеристиках профессий и потребности в них на рынке труда, предусматривающей возможность введения дополнительных местных данных. Первоначально рассчитанная на использование студентами колледжей и университетов, *SIGI-PLUS* была усовершенствована, и теперь ею могут пользоваться все желающие сменить профессию или выйти на рынок труда на разных этапах жизненного пути. Эта программа построена таким образом, чтобы служить клиенту проводником при изучении релевантных данных и планомерном продвижении к принятию эффективного решения (M. R. Katz, 1993; Norris, Schott, Shatkin, & Bennett, 1986). Однако даже такая удачно спроектированная автоматизированная система не способна взять на себя целиком функцию принятия решения за конкретного человека. Может потребоваться вмешательство опытного консультанта, чтобы побудить клиента до конца разобраться в своих потребностях и характерных особенностях на разных стадиях развития карьеры (см., например, Tiedeman, 1994).

Оценка готовности к трудовой деятельности. Другой тип инструментов, специально разработанных для применения в профконсультировании, имеет отношение к оценке уровня готовности индивидуума к трудовой деятельности. Это понятие возникло в рамках долгосрочного проекта по изучению трудового пути (Super et al., 1957; Super, & Overstreet, 1960). ***Готовность к трудовой деятельности (career maturity)*** указывает на принятие людьми задач профессионального цикла, соответствующих возрастному уровню, и на эффективность выполнения ими таких задач. Проект под руководством Супера (Super) представлял собой 20-летнее лонгитюдное исследование около 100 мальчиков-девятиклассников. Результаты позволили сделать вывод, что главной задачей профессионального цикла на уровне младшей средней школы являет-

ся подготовка к выбору профессии. Другие исследования профессионального самоопределения и развития, как лонгитюдные, так и выполненные методом поперечных срезов, постепенно способствуют дополнению и расширению нарисованной Супером картины трудового пути человека (см., например, Crites, 1969; Gribbons, & Lohnes, 1982; Super, 1980, 1985). Важным результатом более современных исследований стал вывод о том, что поведение, типичное для разных этапов профессионального цикла, может встречаться на протяжении большей части жизненного цикла человека, даже если на каждой стадии жизни преобладает один из этих типов поведения. Такие перемены обусловлены множеством факторов, вызвавших изменения как в личной жизни людей, так и в характере профессиональной деятельности (Kummerow, 1991; Lowman, 1991, 1993; Pickman, 1994; Walsh, & Osipow, 1993).

Одним из побочных продуктов исследований трудового пути человека стало создание стандартизованных средств измерения готовности к трудовой деятельности (Kapes et al., 1994, p. 241–272). В качестве примера можно привести разработанный Супером Опросник профессионального самоопределения (*Career Development Inventory* [*CDI*]), предназначенный для оценки готовности выбрать профессию и выявления любых аспектов построения карьеры, в которых людям может понадобиться помощь (A. S. Thompson, & Lindeman, 1981, 1984). Инструмент того же типа — Инвентарь мнений о карьере (*Career Beliefs Inventory* [*CBI*] — Krumboltz, 1991), который разрабатывался в качестве вспомогательного средства профконсультантов специально для выявления любых мнений и представлений индивидуума, которые могут блокировать успешное достижение им карьерных целей.

Клиническая оценка

То, что клиницист делает при оценивании клиента, можно рассматривать как особый случай познания человека человеком, или межличностного восприятия — процесса, посредством которого каждый из нас продвигается к узнаванию и пониманию другого (Kruglanski, 1989). В клинической ситуации, однако, точность оценочного суждения отличается от точности обыденных суждений о другом человеке в нескольких существенных отношениях. Много написано о таких функциях клинициста, как обработка, синтез и интерпретация данных. Исследования процесса клинической оценки пролили свет на некоторые из возможных источников ошибок в этом процессе, включая влияние культурных стереотипов и опору на ошибочные принципы предсказания. Примеры последних включают игнорирование базиса (*base rate*) и эффектов регрессии, а также допущение о большей надежности предикторов, имеющих более высокие интеркорреляции (см., например, L. R. Goldberg, 1991).

При условии одинакового набора фактов, таких как тестовые показатели или биографические данные, может ли клиническая оценка обеспечить более точные предсказания последующего поведения по сравнению с теми, которые можно было бы получить обычным путем, применяя уравнения регрессии или другие эмпирические формулы? Этот вопрос имеет как практическое, так и теоретическое значение. Ведь стоит только разработать статистическую методику или алгоритм, как его применение можно передать техническому персоналу или компьютеру, освободив тем самым клинициста для выполнения других функций. В классической книге под названием «Сравнение клинического и статистического предсказания» *(Clinical Versus Statistical Prediction)* Мил (Meehl, 1954) рассмотрел процесс клинической оценки и проанализиро-

вал исследования, сравнивающие два типа предсказания. Мил показал, что, за одним, требующим дополнительной проверки исключением, применение типовых статистических процедур давало, по меньшей мере, столько же верных предсказаний, сколько их давал клинический анализ, а часто — даже больше. Публикация книги Мила и последовавшие за этим подтверждения его вывода, вызвали оживленную, не затихающую до сих пор дискуссию (см., например, Anastasi, 1988b, p. 511–515; Dawes, Faust, & Meehl, 1993; Kleinmuntz, 1990).

Несмотря на видимое превосходство актуарного подхода, при котором клинические и статистические процедуры применяются к одним и тем же данным, важно иметь в виду, что клинический способ предлагает определенные преимущества. Основным вкладом клинического метода, например, является получение данных путем тщательного интервьюирования и наблюдения за поведением в тех областях, где просто отсутствуют удовлетворительные тесты. Кроме того, клинический метод более пригоден, чем статистический, для изучения неповторимых или редких событий, частота которых слишком мала, чтобы можно было реализовать подходящие статистические стратегии.

В общем, при нынешнем уровне наших знаний, наиболее эффективная процедура обычно сочетает клинический и статистический подходы (Matarazzo, 1990). Клиницисту следует использовать все объективные тестовые данные и актуарные методики, применимые в конкретной ситуации, дополняя эту информацию фактами и выводами, которые можно получить только клиническими методами. Валидность клинических прогнозов относительно реальных результатов должна систематически исследоваться всякий раз, когда это возможно.[1] Нужны также дополнительные данные о согласованности прогнозов в отношении одних и тех же людей, сделанных разными клиницистами в одно время и теми же клиницистами, но в разное время. Насколько это возможно, сам процесс составления клинических предсказаний и признаки, на которых они основываются, следует фиксировать в явной форме в клинических протоколах. Такая практика могла бы не только облегчить исследования и обучение, но и способствовала бы росту доверия к надежным данным и оправданным интерпретациям. Наконец, «клиницист как инструмент» (*clinician as instrument*) остается важным понятием, о чем свидетельствуют непрекращающиеся исследования характеристик, снижающих точность клинической оценки. На основании серии продолжающихся исследований Спенглер и его коллеги пришли к выводу, что психологи с относительно низким уровнем когнитивной сложности более склонны к тенденциозным клиническим оценкам, чем психологи с более высокой когнитивной сложностью (Spengler, & Strohmer, 1994; Walker, & Spengler, 1995).

Акт оценки: завершающий синтез. В главе 18 будут рассмотрены некоторые из общих проблем, касающихся сообщения результатов теста, в особенности этических и социальных последствий. Для клинициста такое сообщение обычно включает подго-

[1] В настоящее время планируются несколько крупномасштабных проектов такого рода, имеющих целью определение валидности решений или предсказаний, основанных на результатах применения методик оценки. Одно такое исследование, уже начатое при содействии Общества оценки личности (*Society for Personality Assessment*), представляет собой полный литературный обзор (за которым последует метаанализ) исследований, использовавших инструменты или процедуры оценки личности для составления прогнозов в отношении различных аспектов терапии или медицинских, правовых и бытовых последствий (Handler, & Meyer, 1996, Spring/Summer).

товку письменного отчета о результатах теста или же медицинского заключения, которые могут обсуждаться, а могут и не обсуждаться с клиентом, его родителями, учителями или с другими специалистами. Даже в случаях, когда письменного заключения не требуется, лучше его все же подготовить, поскольку оно может быть использовано в качестве документа для последующих ссылок. К тому же написание заключения помогает клиницисту упорядочить и прояснить собственные мысли по поводу конкретного случая и отточить свои интерпретации. Письменное заключение представляет собой завершающую фазу синтезирования данных клиницистом. Что касается его содержания, в заключении должны быть использованы все доступные клиницисту источники данных (как тестовых, так и нетестовых).

Целый ряд книг содержит методические указания по написанию отчетов и заключений.[1] Не повторяя многочисленные советы, которые можно найти в таких источниках, мы сосредоточимся на некоторых важных моментах. Прежде всего следует сказать, что не существует единой стандартной формы или схемы для всех заключений. Как содержание, так и стиль заключения могут и должны меняться в зависимости от цели оценки, ситуации, в которой она производится, адресата заключения, теоретической ориентации и специализации клинициста. Особенно важно, чтобы заключение соответствовало потребностям, интересам и уровню подготовки тех, кто его получит. Например, заключение, адресованное адвокату, будет существенно отличаться от заключения, адресованного психотерапевту. Тем не менее в обоих случаях клиницист должен отобрать из массы собранных им данных те, которые относятся к ответу на вопросы, поставленные с самого начала.

В заключении в первую очередь должны быть отражены отличительные особенности индивидуума, т. е. черты с высокими и низкими оценками, а не черты, по которым он имел показатели, близкие к средним. Проверка эффективности заключения состоит в том, чтобы посмотреть, написано ли оно исключительно для данного человека, или в той же мере применимо к другим людям. Составить псевдозаключение из общих стереотипных утверждений, применимых к большинству людей, относительно легко. Значительное количество исследований показало, что подобные заключения большинство людей охотно признают как «удивительно точные» описания их личности (Goodyear, 1990; Klopfer, 1983; Snyder, & Larson, 1972; Tallent, 1992, pp. 236–238). Такая псевдовалидизация была названа «эффектом Барнума», по имени Финеаса Т. Барнума, известного шоумена, которому приписывают высказывание «каждую минуту на свет появляется простофиля». На доверии к общим описаниям личности строится деятельность предсказателей судьбы и других шарлатанов.

Передний план заключения должен отводиться под интерпретации и выводы, хотя в некоторых случаях протоколы тестов и другие подробные данные могут прилагаться к нему отдельно.[2] Конкретные данные, такие как характерные реакции и показатели субтестов, обычно приводятся только для иллюстрации или пояснения какой-либо

---

[1] Сэттлер (Sattler, 1988, chap. 23) дает исчерпывающий обзор встречающихся при этом затруднений и ловушек, вместе со множеством советов по их преодолению, и проводит краткий анализ исследований, касающихся написания заключений. Другими полезными источниками, с примерами клинических заключений, являются работы Ownby (1991) и Tallent (1993).

[2] См. июньский номер журнала *American Psychologist* (June 1996), где опубликовано «Положение о раскрытии тестовых данных», подготовленное Комитетом по психологическим тестам и психологической оценке АПА (*Committee on Psychological Tests and Assessment of the American Psychological Association*).

особенности. Текст заключения должен быть тщательно организованным и связным. Его следует писать просто, преследуя цель сообщить получателю важную информацию, а не сбить его с толку. Пособия по подготовке актов оценки (*assessment reports*) обычно содержат полезные советы по написанию заключений, также как и ссылки на общепринятые руководства по стилистике. Маленькая и крайне занимательная книжка Странка и Уайта (Strunk & White, 1979) «Азы стиля» (*The Elements of Style*) наверняка облегчит задачу составителю заключения и избавит от головной боли его потенциального читателя.

Роль компьютеров в психологической оценке

Такая услуга, как автоматизированный подсчет показателей многих видов тестов, включая личностные опросники типа стандартизованных самоотчетов, доступна пользователям уже в течение нескольких десятилетий. Большинство компьютерных программ выполняют также типовой статистический анализ, предоставляя такую дополнительную информацию, как разные типы производных показателей, интервалы показателей в единицах *SEM* и профили показателей. Описательные заключения (носящие рекомендательный характер!) представляют собой более тонкое применение машинной технологии к составлению отчетов и использованию тестовых данных (Butcher, 1987; Moreland, 1992). По существу, эти программы работают с большими базами данных в виде качественных интерпретирующих утверждений, привязанных к определенным уровням или паттернам количественных показателей. Помимо экономии времени клинициста, эта процедура обладает и другими преимуществами. Компьютер может методично и последовательно вести поиск информации в гораздо более обширной базе данных, чем это мог бы сделать любой клиницист в одиночку; он также способен безошибочно применять к актуарным данным более сложные правила отбора интерпретирующих утверждений и, наконец, он позволяет вводить в процесс подготовки заключения другие уместные переменные, такие как демографические данные из разных нормативных совокупностей.[1]

Бесспорно, потенциальный вклад компьютеров в область психологической оценки выглядит впечатляюще (Butcher, 1987; Eyde, 1987; Gutkin, & Wise, 1991; Moreland, 1992). Однако большая часть этого потенциала только начинает осваиваться (Embretson, 1992). Например, применение ветвящихся процедур (*branching techniques*) и адаптивного тестирования, преимущества которых сейчас широко признаны в области тестирования способностей, только начинает проникать в сферу тестирования личности (Ben-Porath, & Butcher, 1986; Jackson, 1985, 1991) и пока еще не привело к созданию инструментов, пригодных для клинического использования.

С другой стороны, компьютерные технологии привели к созданию и быстрому распространению множества новых инструментов для оценки когнитивного функционирования, которые уже используются в клинической нейропсихологии, а также в

[1] Большинство издателей и дистрибьюторов программного обеспечения для построения профилей тестовых показателей, составления описательных заключений и других вариантов машинной интерпретации тестов снабжают потенциальных пользователей образцами заключений и демонстрационными программами. В добавление к этому, такое издание, как *Psychware Sourcebook* (Krug, 1993), являющееся одним из наиболее полных источников текущей информации о программных продуктах в области автоматизированной психологической оценки, содержит репродукции множества образцов отчетов и заключений.

области оценки недостаточной специфической обучаемости и расстройств внимания (см., например, Krug, 1988, 1993; Stoloff, & Couch, 1992). В обозримом будущем вряд ли стоит надеяться на то, что адекватную оценку нейрокогнитивного функционирования можно будет получить исключительно средствами компьютеризованного тестирования (Golden, 1987). Тем не менее компьютеры уже сейчас делают возможным более точное варьирование условий предъявления задачи испытуемому, с тем чтобы оценить выполнение ее различных компонентов. Кроме того, они позволяют регистрировать и оценивать параметры реакций (например, распределение ответов во времени) способами, применение которых невозможно при проведении бланковых или даже индивидуальных тестов. «МикроКог: Оценка когнитивного функционирования» (*MicroCog: Assessment of Cognitive Functioning* — Powell et al., 1993; Powell, & Whitla, 1994a, 1994b) — пример недавно разработанной компьютеризованной батареи, предназначенной для скрининга по возможным признакам нарушения когнитивных функций у взрослых людей. «МикроКог» состоит из 18 субтестов в таких областях, как внешнее и внутреннее внимание, память, рассуждение и вычисление, оперирование пространственными образами и время реакции, разработанных с прицелом на максимальное использование уникальных возможностей компьютерных технологий. Потенциальные области применения этого относительно быстрого и недорогого теста включают оценку когнитивного снижения у пожилых людей (в диапазоне от легкого до умеренного уровней) и мониторинг познавательной деятельности работников, которые могут подвергаться вредным для организма воздействиям; его также можно использовать и для других целей, прежде всего там, где требуется точная оценка нейрокогнитивных изменений.

Было разработано много других новых инструментов для оценки отдельных функций, и, несомненно, их станет еще больше в ближайшем будущем. Среди самых первых тестов этой группы есть несколько тестов непрерывной деятельности (*continuous performance tests*), таких как Тест параметров внимания (*Test of Variables of Attention* [*TOVA*] — Leark, Dupuy, Greenberg, Corman, & Kindschi, 1996), который доступен пользователям как в слуховом, так и в зрительном вариантах. К дополнительным примерам относятся Тест непрерывной следящей деятельности (*Vigil Continuous Performance Test* [*VIGIL*] — Cegalis, Cegalis, Bowlin, 1993) и Слуховой тест последовательного сложения в заданном темпе (*Paced Auditory Serial Addition Test* [*PASAT*] — Cegalis & Birdsall, 1995); оба содержат мультимедийные комплекты программного обеспечения, позволяющие оценивать внимание в зрительной и слуховой модальностях. На рис. 17–5 изображен ребенок, проходящий зрительный вариант Теста параметров внимания (*TOVA*).

Если посмотреть с другой стороны, то компьютеры обладают еще и потенциальными возможностями в плане интеграции данных из многих источников, включая все виды имеющихся тестов, истории болезни и данные непосредственного наблюдения за поведением (см., например, Watkins, & McDermott, 1991). Хотя, вероятно, и можно было бы предать компьютеру выполняемую клиницистом функцию синтезирования данных по индивидуальному случаю, но необходимая для разработки и поддержки таких интегративных программ компьютеризованная база данных пока еще отсутствует.

Нужно иметь в виду, что в этих современных приложениях тестирования встречается немало подводных камней (Moreland, 1992). Большинство доступных пользователям систем машинной интерпретации тестов объединяют в себе клинические и статистические методы. Специфическая «смесь» количественных данных и клинических оценочных суждений различается от системы к системе, так же как различается



Рис. 17–5. Ребенок, проходящий Тест параметров внимания (*TOVA*), отвечает на стимулы, нажимая на внешний миниатюрный выключатель, подсоединенный к одному из портов компьютера
(С любезного разрешения American Guidance Service)

техническое качество баз данных и клиническое качество суждений. Кроме того, информация, необходимая для оценки конкретной системы, часто оказывается недоступной вследствие мер по охране собственности фирмы-производителя.

Скорее всего, именно отсутствие необходимой технической информации вызвало широкую обеспокоенность возможным неправильным использованием программ машинной интерпретации тестов (Eyde, & Kowal, 1987; Fowler, & Butcher, 1986; Matarazzo, 1986 a, 1986 b; Moreland, 1987). Некоторые программы уже отвечают соответствующим научным и профессиональным стандартам или перерабатываются в настоящее время для достижения этой цели; сведений о качестве ряда программ вообще нет, они никогда не рецензировались квалифицированными специалистами; и несравнимо большее их число явно переоценивается вследствие непроверенных заявлений в изданиях рекламного характера. Первые инструктивные материалы по оценке и использованию услуг машинной интерпретации были опубликованы Американской психологической ассоциацией в 1986 г.; переработанные положения этих инструкций и дополнения к ним включаются в новые *Стандарты тестирования*, работа над которыми ведется в настоящее время. Некоторые дополнительные инструктивные материалы по использованию компьютеризованных средств психологического тестирования можно найти в других публикациях (см., например, Bersoff, & Hofer, 1991; Moreland, 1992).

Заключительные замечания. В целом, область психологической оценки в том виде, как ее применяют на практике разные специалисты, подвергается столь же быстрым изменениям, как и другие области, рассмотренные в этой главе. В добавление к непрерывно расширяющемуся множеству разработок с использованием вычислительных машин и к другим тенденциям, которые уже выделились к концу последнего

раздела данной главы, отмечается возобновившаяся потребность в средствах оценки, ориентированных на положительный полюс психического здоровья, а не на психопатологию. Один недавний пример — Опросник качества жизни (*Quality of Life Inventory* [*QOLI*] — Frisch, 1994), который служит мерой удовлетворенности жизнью и может применяться как при планировании лечения, так и оценке его результатов. Несколько других инструментов этого типа сейчас находятся на разных стадиях разработки. Еще одной интересной и пока единственной новинкой такого рода является Опросник кросс-культурной адаптивности (*Cross-Cultural Adaptability Inventory* [*CCAI*] — C. Kelley & Meyers, 1993), который, как можно судить по названию, представляет собой инструмент самооценки, предназначенный помочь людям принять решение о своей готовности приспособиться к другим культурам. Вопросник для оценки адаптации студентов к колледжу (*Student Adaptation to College Questionnaire* [*SACQ*] — R. W. Baker, & Siryk, 1989) — еще один инструмент, который, подобно *CCAI*, служит типичным примером применения психологического тестирования для самопонимания и самосовершенствования. Такое применение тестирования является прямым следствием влияния психологического консультирования и, по всей вероятности, получит большое распространение в будущем.

18 ЭТИЧЕСКИЕ И СОЦИАЛЬНЫЕ АСПЕКТЫ ТЕСТИРОВАНИЯ

Психологов давно волновали вопросы профессиональной этики в связи с применением их специфического инструментария как в научных исследованиях, так и в практической работе. Конкретным примером внимания психологов к этим вопросам является систематическая эмпирическая программа, начатая в начале 1950-х гг. с целью разработки первого официального кодекса профессиональной этики. Это масштабное предприятие имело следствием подготовку системы норм, которая была официально принята Американской психологической ассоциацией (АПА) и впервые опубликована в 1953 г. Эти нормы подвергаются непрерывному пересмотру и уточнению, что отражается в периодической публикации исправленных и дополненных версий. Нынешняя версия — *Этические принципы психологов и Кодекс поведения* (*Ethical Principles of Psychologists and Code of Conduct* — APA, 1992)[1] — включает в себя преамбулу и шесть общих принципов, предназначение которых — воодушевлять психологов на служение *высшим идеалам (highest ideals)* этой профессии. Кроме того, туда входят восемь этических норм, с *обеспеченными правовой санкцией правилами (enforceable rules)* для психологов, работающих в самых разных организациях.

Выполнение *Кодекса этики* обеспечивается действиями Комитета АПА по этике (APA Ethics Committee), занимающегося расследованием жалоб на членов ассоциации и вынесением по ним решений. Правила и порядок работы этого комитета, так же как и годовые отчеты о его деятельности, публикуются в журнале *American Psychologist* — официальном печатном органе АПА (см., например, APA, Ethics Committee, 1995, 1996). Крайне необходимый справочник, предлагающий комментарии и иллюстрации применения этических норм, был подготовлен бывшими членами Комитета по этике, участвовавшими в подготовке самой последней версии *Кодекса этики* (Canter, Bennett, Jones, & Nagy, 1994). Другой полезный сборник исторических и современных материалов по этике в психологии, включая обсуждение нравственных дилемм в различных контекстах, составил недавно Берсов (Bersoff, 1995). Объемистый том

[1] Чтобы как можно шире распространить *Этические принципы психологов и Кодекс поведения* (в дальнейшем — *Кодекс этики*), АПА будет бесплатно высылать одну копию этого документа любому, кто обратится с такой просьбой.

под названием «Этика в психотерапии и консультировании» (*Ethics in Psychotherapy and Counseling* (Pope, & Vasquez, 1991) включает главу по этическим проблемам психологической оценки, содержащую полезные практические советы. Наконец, Вайнером (Weiner, 1995a) недавно была написана полезная глава о том, как предупреждать возникновение этических и правовых проблем при оценке личности.

Девяностые годы свидетельствовали о быстром увеличении количества исков на федеральном и местном уровнях, судебных решений и профессиональных руководств (написанных с разных позиций), направленных на ограничение возможностей психологической практики вообще, и использования психологических тестов в частности. Некоторые из этих событий уже обсуждались в предыдущих главах в связи с частными вопросами тестирования и практической работы психологов. Очень часто сочетанное действие этих запретов и предписаний приводило психолога-практика в замешательство и выливалось в непоследовательность действий и конфликты.

Все больше и больше внимания уделяется поставщикам психологических услуг, вынужденным постоянно балансировать между этическими принципами профессии, правовыми и лицензионными нормами, да еще и институционной политикой организаций, в которых они работают. В соответствии с этим, АПА через свои многочисленные бюро и комитеты постаралась обеспечить руководством и информацией своих членов, осуществляя текущий контроль за соответствующими событиями и распространяя стандарты, инструкции, положения и заявления по вопросам, которые могут представлять проблему для практикующих психологов. *Общие инструкции для поставщиков психологических услуг* (*General Guidelines for Providers of Psychological Services* — APA, 1987a) и *Специальные инструкции по поставкам услуг* (*Specialty Guidelines for the Delivery of Services* [APA, 1981 — в настоящее время пересматриваются]) были изданы и распространены для оказания помощи тем, кто занимается профессионально психологической практикой в разных контекстах. Дополнительное руководство в более узких вопросах обеспечивается другими документами, такими как «Инструкции по оценкам опеки над ребенком в бракоразводном процессе» (*Guidelines for Child Custody Evaluations in Divorce Proceedings*), составленные Комитетом по профессиональной практике и профессиональным стандартам (APA, COPPS, 1994). Еще несколько специальных инструкций упоминаются на протяжении этой главы. Вдобавок ко всему, начиная с конца 1980-х гг., АПА выпускает серийное издание, тома которого содержат сводки и анализ законов каждого штата, затрагивающих интересы специалистов в области психического здоровья (*mental health*); ко времени написания этой книги, было опубликовано уже около дюжины томов, а ряд из них — переиздан в обновленной редакции (см., например, Caudill, & Pope, 1995; Petrila, & Otto, 1995; Shuman, 1990, 1993; Wulach, 1991).

Комитет АПА по психологическим тестам и психологической оценке (*APA Committee on Psychological Tests and Assessment* [*CPTA*]), в частности, занимается рассмотрением проблем, касающихся практики надежного тестирования и оценки, и обеспечением технического консультирования других групп АПА в отношении такой практики. Этим комитетом подготовлен ряд документов, позднее упоминаемых в этой главе, содержащих правила разрешения проблем, связанных с применением тестов. *CPTA* также анализирует работу Объединенного комитета по практике тестирования (*Joint Committee on Testing Practices* [*JCTP*]) — группы, созданной АПА и другими профессиональными организациями, связанными с тестированием. В свою очередь, Объеди-

ненный комитет разработал *Кодекс честной практики тестирования в образовании* (*Code of Fair Testing Practices in Education* (JCTP, 1988) и другие материалы, нацеленные на усовершенствование способа применения тестов и на предотвращение их неправильного использования (см., например, Eyde et al., 1988, 1993). В настоящее время *JCTP* готовит «Положение о правах и обязанностях лиц, проходящих тестирование» (*Rights and Responsibilities of Test Takers*).

В главе 1 были рассмотрены некоторые аспекты роли пользователей тестов и то, как выглядит тестирование с позиций тестируемых. В этой главе мы обратимся к этическим и социальным проблемам, влияющим на использование тестов в разных сферах жизни и деятельности. В добавление к вопросам, касающимся профессиональной компетентности, мы кратко обсудим сферу ответственности издателей тестов, право тестируемого на неприкосновенность личной жизни, проблему конфиденциальности и тестирование лиц, принадлежащих к разного рода меньшинствам. Хотя мы и будем в некоторой степени касаться действия законов, детальное рассмотрение множества правовых аспектов практики тестирования выходит за рамки нашей компетенции. По этой причине мы отсылаем заинтересованного читателя к различным литературным источникам, упоминаемым в этой и других главах учебника (см., в частности, главы 9 и 17).

Этические проблемы психологического тестирования и психологической оценки

Начиная с 1970-х гг., во всех областях теоретической и прикладной психологии наблюдалось повышение интереса не только к этическим проблемам, но и к более широким вопросам ценностей (Bersoff, 1995; Diener, & Crandall, 1978; Jacob, & Hartshorne, 1991; Pope, & Vasquez, 1991). В области тестирования, содержательный и наводящий на продуктивные размышления анализ роли ценностей и этических принципов, лежащих в основе разнообразных практических приложений, представлен в работах Eyde, & Quaintance (1988) и Messick (1980b, 1989, 1995). На более конкретном уровне, значительная часть содержания *Кодекса этики* АПА применима к психологическому тестированию. Одна из его норм — «Оценивание, оценка или вмешательство» (*Evaluation, Assessment, or Intervention*) — непосредственно касается разработки и использования методик психологической оценки. Другая норма — «Судебная деятельность» (*Forensic Activities*) — включает раздел, полностью посвященный оценкам в судах разного уровня. В дополнение к этому, этическая норма «Неприкосновенность личной жизни и конфиденциальность» (*Privacy and Confidentiality*), хотя и относится к более широкому содержанию, высоко значима для тестирования, как, впрочем, большинство других общих принципов и ряд других этических норм кодекса (APA, 1992). Некоторые из вопросов, рассматриваемых в *Кодексе этики*, тесно связаны с кругом вопросов, охватываемых *Стандартами тестирования*, о которых шла речь в главе 1. Фактически, содержание самих *Стандартов тестирования* помогает определить границы профессионально ответственного использования тестов.

Помимо АПА, другие родственные профессиональные группы и ассоциации разработали свои этические кодексы и руководства. Среди этих документов, наиболее значимым, с точки зрения тестирования, является «Положение об обязанностях

пользователей стандартизованных тестов» *(Responsibilities of Users of Standardized Tests [«RUST» Statement])*, принятое в 1989 г. Американской ассоциацией консультирования (*American Counseling Association* [*ACA*]). Другой полезный документ — Принципы валидизации и использования методик отбора персонала *(Principles for the Validation and Use of Personnel Selection Procedures)*, разработанные Обществом промышленной и организационной психологии (*Society for Industrial and Organizational Psychology* [SIOP — 1987]) для более узких целей (см. главу 17).

Важным событием для уяснения места тестирования в современном обществе стала публикация книги «Тестирование способностей: области применения, последствия и спорные вопросы» (*Ability Testing: Uses, Consequences, and Controversies* — Wigdor, & Garner, 1982). Этот двухтомный труд представляет собой итоговый отчет о выполнении четырехлетнего проекта, посвященного изучению применения стандартизованных тестов способностей в школах, а также при приеме в высшие учебные заведения и на работу. Начатый в то время, когда проходила широкая публичная дискуссия о значении тестирования, этот проект выполнялся под руководством междисциплинарного комитета при содействии Национального научно-исследовательского совета (*National Research Council*). С начала 1980-х гг. был опубликован ряд других важных исследований и отчетов, касающихся проблемных областей тестирования (см., например, Hartigan, & Wigdor, 1989; Office of Technology Assessment, 1992). В общем, результаты этих различных исследовательских групп еще раз подтвердили обоснованные и часто повторяемые выводы о полезности и возможных нарушениях правил использования тестов способностей.

Принимающая все более широкие масштабы заинтересованность правительственных органов и организаций в применении психологических тестов и других средств оценки привела к созданию Совета по тестированию и оценке (*Board on Testing and Assessment* [*BoTA*]), который был учрежден в 1993 г. при поддержке министерств обороны, просвещения и труда (см. приложение Б). Совет по тестированию и оценке продолжает деятельность Национального научно-исследовательского совета. Его главная цель — помочь разработчикам политических стратегий понять и оценить тесты и другие средства оценки, используемые как орудия государственной политики. Совет сосредоточен на рассмотрении актуальных вопросов тестирования и оценки в самых разных сферах жизнедеятельности и уже опубликовал ряд отчетов по таким темам, как «Цели законодательной инициативы в сфере образования–2000» (Feuer, & Kober, 1995), «План усовершенствования Батареи тестов общих способностей» (BoTA, 1995) и «Оценка изменений характера работы и их последствий для системы образования» (Black, Feuer, Guidroz, & Lesgold, 1996).

Оценка квалификации пользователей и профессиональная компетентность

Входящий в Кодекс этики принцип компетентности гласит, что психологи «предоставляют только те услуги и используют только те методики, которые соответствуют их квалификации, обеспечиваемой образованием, специальной подготовкой или опытом» (APA, 1992, p. 1599).Что касается тестов, требование, чтобы они использовались только специалистами с соответствующей подготовкой, является первым шагом к за-

щите тестируемых от неправильного использования тестов.[1] Конечно, требуемая квалификация меняется в зависимости от типа теста. Так, для правильного применения индивидуальных тестов интеллекта и большинства личностных тестов требуется относительно длительный период интенсивного обучения и работы под наблюдением наставника, в то время как для тестирования учебных достижений или профессиональной умелости нужна гораздо менее специализированная психологическая подготовка. Следует также заметить, что студенты, которые проходят тесты с учебными целями в специальных курсах, обычно не готовы к самостоятельному проведению тестов с другими людьми или к правильной интерпретации тестовых показателей.

Хорошо подготовленные пользователи выбирают тесты, оптимально подходящие как для той цели, с которой они проводят тестирование, так и для тех лиц, которых они тестирует. Они также знакомы с научной литературой по выбранному тесту и способны оценить его технические достоинства относительно таких характеристик, как нормы, надежность и валидность. При проведении теста они чувствительны к условиям и обстоятельствам, могущим влиять на его выполнение, в частности к упоминавшимся в главе 1. Они делают выводы или дают рекомендации только после рассмотрения тестового показателя (или показателей) в контексте другой релевантной информации об индивидууме. Главное же, они должны быть достаточно осведомленными в науке о человеческом поведении, чтобы избежать неоправданных выводов в своих интерпретациях результатов тестирования. В тех случаях, когда тесты проводятся младшим психологическим персоналом или ассистентами, либо лицами без должной теоретической подготовки в области психометрии и без достаточного практического опыта, для обеспечения необходимых условий правильной интерпретации выполнения теста важно, чтобы их по крайней мере консультировал психолог, обладающий соответствующей квалификацией.

Кого считать квалифицированным психологом? Очевидно, из-за разнообразия областей психологии и, следовательно, специализации в подготовке ни один психолог не может быть одинаково сведущим во всех областях, даже внутри более узкой сферы психологического тестирования и оценки (см. главу 17). Как признание этого факта, *Кодекс этики* призывает психологов «признавать границы своей компетентности и ограниченность своего профессионального опыта» (APA, 1992, p. 1599). Следствия этого этического обязательства раскрываются в цитированном выше принципе компетентности.

Значительным шагом, повысившим профессиональные стандарты и давшим общественности критерий для определения уровня квалифицированности психолога, было принятие штатами законов о лицензиях и аттестации психологов. В настоящее время все штаты и округ Колумбия имеют такие законы; на большей части территории Канады также приняты законы, регулирующие психологическую практику (что касается сводки всех этих законов, см. APA, 1993, p. XLII–XLV). Хотя термины «лицензирование» и «аттестация» часто используются как взаимозаменяемые, в психологии аттестация обычно указывает на юридическую защиту названия профессии «психолог», в то время как лицензирование относится к регулированию психологи-

---

[1] Обсуждению роли пользователя тестов, а также проектов по оценке квалификации пользователей и их специальному обучению, реализуемых рабочими группами *JCTP* (Eyde et al., 1988, 1993; Moreland et al., 1995), посвящена часть главы 1. Канадская психологическая ассоциация и Британское психологическое общество также предприняли определенные шаги в создании своих систем для определения квалификации пользователей тестов (D. C. Brown, 1995).

ческой практики независимо от того, как называется профессия лица, предоставляющего психологические услуги. Поэтому законы о лицензиях должны включать определение психологической практики. Большинство штатов начинали с более простых законов об аттестации и постепенно продвигались к принятию законов о лицензировании. Но и в том и в другом случае законом обычно требуется докторская степень по психологии, определенный стаж работы под руководством наставника и удовлетворительная сдача квалификационного экзамена. Законы о лицензировании, как правило, включают перечень оснований для применения к психологам административных мер воздействия, от штрафов и взысканий до приостановки действия и аннулирования лицензии.

На более высоком уровне профессиональная аттестация проводится американским Советом по профессиональной психологии (АВРР — см. приложение Б). При условии высокого уровня профессиональной подготовки и опыта работы по обозначенной специальности Совет выдает аттестаты в таких областях деятельности, как клиническая психология, психологическое консультирование, промышленная/организационная психология и школьная психология, наряду с другими, по которым есть отдельные профессиональные советы. В текущем каталоге Американской психологической ассоциации перечислены все выданные аттестаты по каждой специальности, их перечень можно также получить, запросив его в самом Совете по профессиональной психологии. Решения Совета как частным образом организованного объединения отделов в рамках общей профессии не имеют той обязательной силы, которая есть у агентств, проводящих лицензирование и осуществляющих аттестацию на основе законов штата.

Произошедшие в прошлом десятилетии изменения в системе здравоохранения и на рынке профессиональных услуг придали особую остроту вопросу о доверии к выдаваемым аттестатам и лицензиям, позволяющим заниматься психологической практикой. Поэтому Американская психологическая ассоциация предприняла ряд шагов для обеспечения мирного урегулирования многих потенциальных конфликтов, которые заложены в сложившейся к настоящему времени обстановке. Одним из этих шагов является создание Колледжа профессиональной психологии АПА (Sleek, 1995), который выдает свидетельства о присвоении разнообразных квалификаций внутри психологии на основе процедуры, учитывающей результаты экзаменов, а также необходимые требования к образованию и опыту работы. Другим шагом стала разработка процесса, посредством которого можно добиться официального признания специальностей и квалификаций в области психологической практики (APA, Joint Interim Committee for the Identification and Recognition of Specialties and Proficiencies, 1995 a, 1995b). Несомненно, в ближайшем будущем появится руководство по такому частному вопросу, как определение квалификации пользователей тестов, а также будут разработаны процедуры выдачи свидетельств специалистам по психологической оценке.

Профессиональная ответственность издателей тестов

Право на приобретение тестов обычно предоставляется лицам, соответствующим определенному квалификационному минимуму. В каталогах основных издателей тестов приводятся требования, которым должны удовлетворять их покупатели. Некоторые издатели распределяют тесты по уровням с точки зрения требований к квалифи-

кации возможных пользователей, начиная с тестов учебных достижений и тестов профессионального мастерства и кончая такими клиническими инструментами, как индивидуальные тесты интеллекта и большинство личностных тестов. Кроме того, делаются различия между индивидуальными и уполномоченными конкретной организацией покупателями одних и тех же тестов. Студенты-дипломники и аспиранты, которым может понадобиться определенный тест для выполнения практических заданий или проведения исследований, должны иметь заявку на приобретение, подписанную еще и их преподавателем психологии, берущим на себя ответственность за надлежащее использование данного теста.[1]

Меры по ограничению распространения тестов преследуют двоякую цель: неразглашение тестовых материалов и предупреждение их неправильного применения. Следует, однако, отметить, что хотя распространители тестов могут прилагать все усилия для достижения этой цели, контроль, который они в состоянии осуществить, неизбежно ограничен. В некоторых случаях у издателей просто нет возможности проверить свои сомнения в отношении квалификации покупателей тестов (см., например, Oles, & Davis, 1977). Кроме того, официальные квалификационные документы обеспечивают лишь самый грубый отсев. Очевидно, например, что степень магистра или даже доктора психологии, лицензия штата и аттестат Совета по профессиональной психологии не обязательно означают, что данный человек достаточно квалифицирован для использования какого-то конкретного теста, или что его подготовки достаточно для правильной интерпретации его результатов. Основная ответственность за надлежащее использование тестов, в конечном счете, возлагается на индивидуального пользователя или заинтересованную организацию.

Еще одна сфера профессиональной ответственности связана с маркетингом психологических тестов, осуществляемым авторами и издателями. Не следует выпускать для общего применения неподготовленные тесты. Недопустимы какие-либо заявления о достоинствах теста при отсутствии достаточно объективных оснований. Когда недоработанный тест распространяется только с исследовательскими целями, это условие должно быть четко оговорено, а распространение теста соответственно ограничено. В руководстве к тесту должны приводиться все необходимые и достаточные данные для оценки самого теста и полная информация о его проведении, подсчете показателей и нормах. Руководство должно давать фактическое представление о том, что известно о тесте, а не быть средством его рекламы, представляющим тест в выгодном свете. Обязанностью авторов и издателей является достаточно частый пересмотр тестов и их норм, с тем чтобы предупредить их старение. Разумеется, время, за которое тест устаревает, весьма различно и зависит от природы теста.

Тесты, которые должны быть закрытыми вследствие их применения при отборе и расстановке кадров или при принятии диагностических решений, по понятным причинам не могут публиковаться в средствах массовой информации ни целиком, ни частично. Предание гласности даже каких-либо отдельных заданий теста может сделать невалидным последующее применение теста к другим людям. Вдобавок ко всему, публикация тестов в широкой печати может привести к формированию психологи-

[1] Комитет АПА по психологическим тестам и психологической оценке (APA's Committee on Psychological Tests and Assessment, 1995) подготовил специальное Положение по использованию закрытых психологических тестов в обучении аспирантов и студентов последних курсов психологических отделений и факультетов.

чески вредящих самооценок у некоторых читателей, когда результаты самопроверки противоречат сложившимся представлениям о себе. Еще одним непрофессиональным (за редким исключением) использованием психологических тестов является тестирование по почте. Такой способ не только не обеспечивает контроля условий тестирования, но обычно предполагает интерпретацию тестовых показателей без привлечения другой важной информации о тестируемом. Если вынести за скобки те редкие исключения, когда инвентари интересов или ценностей рассылаются по почте достаточно подготовленным и высоко мотивированным лицам, результаты тестирования в таких условиях могут оказаться не только бесполезными, но и вредящими.[1]

Начиная с 1980-х гг., издатели тестов начали предпринимать некоторые шаги для обеспечения того, чтобы издаваемые и распространяемые ими тесты использовались надлежащим образом, а их показатели корректно интерпретировались. С этой целью они попытались расширить и укрепить связи со своими клиентами, а также улучшить понимание роли тестирования среди широких слоев населения. Большинство издателей приняли участие вместе с АПА и другими национальными организациями в проектах Объединенного комитета по практическому применению тестирования (*Joint Committee on Testing Practices* [*JCTP*]), посвященных разработке квалификационных требований к пользователям тестов и программ специальной подготовки пользователей (Eyde et al., 1988, 1993). В добавление к этому они учредили Ассоциацию издателей тестов (*Association of Test Publishers* [*ATP*] — см. приложение Б). Члены этой организации взяли на себя обязательство следовать принципу честности в распространении своей продукции и предоставлении услуг по психологической оценке, а также содействовать повышению их ценности в глазах общественности. *ATP* недавно опубликовала второй выпуск серии Типовых инструкций по обеспечению честности тестирования при приеме на работу (*Model Guidelines for Preemployment Integrity Testing* — ATP, 1996).

Защита неприкосновенности личной жизни

В связи с тестами, особенно личностными, возникает вопрос о вторжении в личную жизнь. В докладе, озаглавленном «Неприкосновенность личной жизни и поведенческие исследования» (*Privacy and Behavioral Research*, 1967), право на неприкосновенность личной жизни определяется как право индивидуума самостоятельно решать, в какой степени ему делиться своими мыслями, чувствами и событиями личной жизни с другими людьми, причем это право далее характеризуется как «совершенно необходимое для обеспечения свободы и самоопределения» (р. 2). Поскольку некоторые тесты эмоциональных, мотивационных и аттитюдных характеристик личности обязательно имеют замаскированный характер, то обследуемый человек может раскрывать в процессе такого теста эти характеристики, не отдавая себе в этом отчета. Для эффективности тестирования иногда приходится скрывать от тестируемого специфику интерпретации его ответов. Тем не менее никто не должен подвергаться какому бы то ни было тестированию под выдуманным предлогом. В этой связи первейшая

[1] Комитет по этике АПА недавно подготовил официальное заявление по поводу политики «прохождения тестов на дому» в ответ на запрос, касающийся правомерности рассылки *MMPI* для заполнения в домашних условиях (APA, Ethics Committee, 1994, р. 665–666).

обязанность тестирующего — довести до сознания тестируемых будущее использование результатов теста.

Хотя озабоченность вторжением в личную жизнь выражалась, в основном, в связи с личностными тестами, логично распространить ее на все типы тестов. Бесспорно, любой из тестов интеллекта, способностей или достижений может выявить такие пробелы в навыках и знаниях, которые тестируемый предпочел бы скрыть. Более того, простое наблюдение за поведением человека во время целенаправленной беседы, случайного разговора или встречи может открыть такую информацию о нем, которую он не хотел бы выдавать и обнаружил невольно. То, что именно психологические тесты часто выбирались на роль главных обвиняемых в дискуссиях о посягательстве на тайну личной жизни, вероятно, отражает преобладание ошибочных представлений о сути тестов, равно как и злоупотребление ими в качестве единственной основы для принятия решения в отношении конкретных людей. Если бы все тесты понимались как средства измерения выборочных образцов поведения, не обладающие никакими мистическими способностями проникать за границы поведения, то распространенные опасения и подозрения быстро бы рассеялись. Аналогично, если бы тесты интерпретировались в контексте комплексных оценок всякий раз, когда дело касается важных для конкретного человека решений, снизилось бы и то чрезмерное значение, которое часто придается полученному результату по какому-либо тесту.

Следует заметить, что всякое исследование поведения, независимо от того, используются ли в нем тесты, или другие обсервационные методы, заключает в себе возможность вторжения в личную жизнь. Разумеется, как ученые, психологи преследуют цель преумножения знаний о человеческом поведении. Возникающие при этом ценностные конфликты каждый раз должны разрешаться с учетом конкретной ситуации.[1] Проблема, очевидно, не столь проста и потому стала предметом широкого обсуждения.[2] Обеспечение тайны личности никакие универсальные правила гарантировать не могут, они только служат общими ориентирами, но в конкретном случае эти ориентиры не в состоянии заменить этическую сознательность и профессиональную ответственность самого психолога, принимающего свои решения в соответствии с частными обстоятельствами.

Одним из релевантных факторов является цель проведения тестирования — индивидуальное консультирование, отбор и распределение персонала или научное исследование. Например, в условиях клиники и при индивидуальной консультации клиенты обычно более охотно раскрываются с тем, чтобы получить помощь в решении своих проблем. Однако, какой бы ни была цель тестирования, защита неприкосновенности личной жизни связана с двумя ключевыми понятиями: релевантности (*relevance*) и осведомленного согласия (*informed consent*). Запрашиваемые у индивидуума сведения должны быть *релевантны* заявленным целям тестирования. Важное следствие этого принципа состоит в том, что должны быть использованы все реальные возможности, чтобы установить валидность тестов в отношении той диагностической или

---

[1] Такой документ, как Этические принципы проведения исследований на людях (*Ethical Principles in the Conduct of Research with Human Participants* — APA, 1982), обеспечивает некоторые ориентиры в этом отношении.

[2] См., например, книгу Ф. Аллана Хэнсона (Hanson, 1993), посвященную критике тестирования и его роли в современном обществе. Хотя его трактовка проблемы явно базируется на идеологии, несовместимой с тестированием, и далека от беспристрастной, она может представлять интерес для читателей в силу выбора автором антропологической перспективы.

прогностической цели, ради которой они используются. Недавние события в юридической жизни Америки, такие как слушание дела *Soroka v. Dayton Hudson* (см., например, Merenda, 1995) и Закон об инвалидах-американцах от 1990 г. (ADA – P.L. 101–336), подчеркнули важность сведения содержания анкетирования и собеседования при приеме на работу к необходимому минимуму и обеспечения доказуемой релевантности собираемой информации о работнике в случае оценки выполнения работы (см., например, Bruyere, & O'Keeffe, 1994; D. C. Brown, 1996; Herman, 1994, chap. 2). В деле *Soroka v. Dayton Hudson* претенденты на получение работы оспорили использование скрининг-теста на том основании, что его вопросы о религиозных убеждениях и сексуальных предпочтениях (взятые из *MMPI* и *CPI*) являются вторжением в личную жизнь и носят дискриминационный характер. Хотя это дело было улажено без вынесения окончательного судебного решения, некоторые разработчики тестов, — включая авторов последних версий *MMPI* и *CPI*, — уже исключили такие пункты из опросников для самоотчета (см. главы 13 и 17).

Понятие *осведомленного согласия* также нуждается в прояснении, а его применение в отдельных случаях требует проявления немалой мудрости (AERA, APA, NCME, 1985). Хотя нынешний *Кодекс этики* содержит явную норму, требующую информированного согласия только в отношении терапии, в неявной форме такое требование присутствует в нормах, касающихся оценки и диагностики в профессиональной сфере, а также в некоторых других частях этого кодекса. Кроме того, разнообразные нормативные документы, принятые на уровне штата, прецедентное право, инструкции лечебных, исправительных и других учреждений, а также распространенные стандарты психологической практики обычно требуют получения информированного согласия как в случаях оценки, так и при реализации программ вмешательства (Canter et al., 1994, p. 67). Конечно, тестируемого следует уведомить о цели тестирования, типе собираемых данных и о том, как будут использованы тестовые показатели. Это, однако, не означает, что ему заранее будут показаны тестовые задания или сообщено, как будут оцениваться его ответы. Не следует также показывать тестовые задания родителям, если обследуются несовершеннолетние.[1] Такая информация обычно делает тест невалидным. Эти и другие специальные вопросы, которые могут возникать по поводу информированного согласия и связанных с ним спорных моментов в ситуациях тестирования и психологической оценки, рассматриваются в главе *Стандартов тестирования*, посвященной правам тестируемых.

Конфиденциальность

Так же как и защита неприкосновенности личной жизни, связанная с ней проблема конфиденциальности данных теста является многоаспектной. Основной ее вопрос: «Кто будет иметь доступ к тестовым результатам?» Ответ на него в конкретных ситуациях определяется рядом соображений, к числу которых относятся защита содержания теста от разглашения, избежание риска неправильного толкования показателей теста и потребность разных лиц в знании результатов тестирования.

[1] Что касается правил получения согласия на оценку и других этических и юридических вопросов психологического оценивания несовершеннолетних, см. Kamphaus, & Frick (1996, chap. 4).

Растущее понимание права человека на получение доступа к данным своего тестирования, а также осознание, что он должен иметь возможность комментировать содержание своего ответа и, в случае необходимости, пояснить или исправить фактическую информацию, заставляет консультантов все больше делать клиента активным участником собственного обследования. Для этих целей тестовые результаты должны быть представлены в удобной для понимания форме, свободной от специальных терминов или профессионализмов и ориентированной на непосредственные задачи тестирования. Против неправильного использования и неверной интерпретации тестовых данных должны быть приняты соответствующие меры предосторожности.

В центре дискуссий по поводу конфиденциальности тестовых данных обычно оказывается их доступность *третьему* лицу, а не тестируемому (или его родителям в случае несовершеннолетия тестируемого) и тестирующему. Основной принцип состоит в том, что такие данные не должны раскрываться без ведома и согласия испытуемого, если их раскрытие не санкционировано решением суда или не разрешено законом для оправданных целей. Когда тесты проводятся в учреждении, например в школе, суде или при поступлении на работу, тестируемого следует проинформировать о целях тестирования, о том, как будут использоваться его результаты, и об их доступности персоналу учреждения, который на законных основаниях будет работать с ними. В тех случаях, когда результаты теста запрашиваются посторонними людьми или другими организациями, например когда возможный наниматель или колледж запрашивают результаты тестирования, проведенного в школе, нужно уже отдельное согласие на передачу данных. Эти же требования применимы и к тестам, проводимым в условиях клиники, при консультировании или с исследовательскими целями. Дополнительные инструкции по этому вопросу можно найти в *Положении о раскрытии тестовых данных* (*Statement on the Disclosure of Test Data* — APA, 1996), разработанном CPTA в помощь психологам, получившим повестки в суд или вызванным в качестве свидетелей и, таким образом, обязанным давать показания, касающиеся сведений о клиенте или тестовых данных, накопленных в ходе практики (APA, COLI, 1996).

Еще одна проблема связана с хранением данных тестирования в учреждениях. С одной стороны, лонгитюдные данные по конкретным людям могут быть весьма полезными не только для исследовательских целей, но и для правильного понимания и консультирования наблюдаемого человека. Но, как это часто бывает, их ценность определяется правильным использованием и верной интерпретацией тестовых результатов. С другой стороны, наличие старых данных открывает путь к такому неверному их использованию, как ошибочные выводы из устаревших результатов и неправомочный доступ к ним с иной, отличной от первоначальной, целью. Было бы явной нелепостью, например, ссылаться на *IQ* или показатель теста овладения чтением, полученные ребенком в 3-м классе, в ходе оценивания абитуриента при приеме в колледж. Аналогично, когда протоколы тестирования хранятся много лет, всегда существует опасность их использования с целями, которых ни сам тестируемый, ни его родители никогда не предполагали и не одобрили бы. Во избежание подобного неправильного обращения с данными, предназначенными для длительного использования либо в интересах наблюдаемого человека, либо с научными целями, доступ к ним должен особенно строго контролироваться. Для учреждений любого типа было бы разумным разработать писаные правила уничтожения, хранения и получения доступа к информации о конкретных людях. Более подробно эта тема затрагивается в изложении Принципов учета и хранения данных (*Record Keeping Guidelines* — APA, COPPS, 1993).

Сообщение результатов теста

В последние годы психологи начали уделять больше внимания сообщению результатов теста в доступной и полезной для получателя форме. Ясно, что подобная информация не может передаваться механически, но должна сопровождаться соответствующими интерпретирующими пояснениями. Общие уровневые характеристики и качественные описания на доступном языке предпочтительней конкретных числовых данных, за исключением тех случаев, когда результаты сообщаются достаточно подготовленным профессионалам. Известно, что даже образованные неспециалисты путают процентили с процентными показателями или с *IQ*, нормы с эталонами и оценки интересов с показателями способностей. Однако более серьезные ошибки интерпретации связаны с выводами, которые делаются на основе тестовых показателей, даже при правильном понимании психометрического смысла последних. Хорошо известным примером служит распространенное убеждение, что *IQ* указывает на устойчивое свойство индивидуума, предопределяющее уровень его интеллектуальных достижений в течение всей жизни.

К числу возможных получателей данных тестирования, кроме самих тестируемых, относятся родители несовершеннолетних, учителя и другие школьные работники, работодатели, психиатры, а также работники судов и исправительных учреждений. При сообщении любой связанной с проведением тестов информации желательно учитывать типичные особенности ее получателя. Это относится не только к общеобразовательному уровню человека и его познаниям в области психологии и тестирования, но и к его ожидаемой эмоциональной реакции на такую информацию. Например, эмоциональная причастность родителей или учителей к жизни ребенка может препятствовать спокойному и разумному принятию фактической информации.

С аналогичными проблемами приходится сталкиваться при сообщении результатов теста самим тестируемым, будь то дети или взрослые.[1] В этом случае применяются те же меры предосторожности против неправильной интерпретации, что и при сообщении данных третьему лицу. В этом отношении *Стандарты тестирования* подчеркивают обязанность тех специалистов, которые применяют тесты в клинической практике и консультировании, давать тестируемым целесообразные и понятные объяснения результатов теста и вытекающие из них рекомендации.

Учитывать эмоциональные реакции на информацию о результатах тестирования особенно важно в тех случаях, когда люди при этом узнают о своих собственных достоинствах и недостатках. Когда кому-то сообщают его тестовые результаты, нужно не только обеспечить их интерпретацию достаточно квалифицированными людьми, но и предоставить возможность получить консультацию специалиста каждому, кого такая информация эмоционально обеспокоит. Например, студент колледжа может быть серьезно обеспокоен, узнав о том, что плохо выполнил тест академических способностей. Одаренный школьник может развить леность и утратить интерес к школьному обучению или утратить интерес к сотрудничеству с другими и стать непослушным,

[1] Сентябрьский выпуск журнала *Psychological Assessment* (September 1992) содержит специальный раздел, посвященный теме обеспечения обратной связи с клиентами, проходящими психологическое тестирование. В особенно полезной статье Поупа (K. S. Pope) рассматриваются десять существенных аспектов обратной связи, которые, по его словам, «возможно, являются самой пренебрегаемой стороной психологической оценки» (1992, p. 265).

если обнаружит, что он гораздо способнее своих сверстников. Такие вредные эффекты могут, конечно, происходить независимо от правильности или неправильности самого тестового показателя. Даже если тест был тщательно проведен, безошибочно обработан и верно интерпретирован, знание результата без возможности обсудить его более подробно может повредить тестируемому.

Консультирующие психологи специально занимались разработкой эффективных способов передачи информации о результатах тестирования своим клиентам (см., например, Hood, & Johnson, 1997, chap. 17). Хотя детальное рассмотрение этого механизма не входит в круг обсуждаемых здесь вопросов, два важных принципа заслуживают особого упоминания. Во-первых, сообщение результатов тестирования должно рассматриваться как неотъемлемая часть сложного процесса консультирования и, соответственно, входить составным элементом в полное отношение «консультант–клиент». Во-вторых, насколько это возможно, консультанты должны привлекать своих клиентов к интерпретации результатов теста в свете поднимаемых ими конкретных вопросов. Важным моментом в консультировании является принятие клиентом предоставляемой ему информации. Специфика ситуации консультирования состоит в том, что если клиент по какой-либо причине отвергает сообщаемую ему информацию, то, по всей вероятности, она полностью бесполезна. С другой стороны, принятие правильно интерпретированных данных теста может иметь терапевтическое значение для клиента, — причем и как сам факт принятия, и как информационное сообщение, — особенно в контексте когнитивно-ориентированной терапии.

Тестирование особых популяций

Текущая обстановка. После 1950-х гг. возросла озабоченность общественности правами этнических групп, женщин, инвалидов и других групп меньшинств.[1] Эта озабоченность нашла отражение в принятии закона о гражданских правах на уровне штатов и на федеральном уровне. В связи с поиском способов улучшения образовательных и профессиональных возможностей меньшинств психологическое тестирование оказалось в центре внимания (Gifford, 1989a, 1989b). Психологическая литература содержит обширные материалы дискуссий по этой теме, результативность которых колебалась от прояснения до окончательного запутывания вопроса. Среди наиболее весомых вкладов в разрешение данной проблемы — ряд меморандумов и руководящих документов, подготовленных профессиональными ассоциациями (см., например, ACA, 1989; APA, Board of Ethnic Minority Affairs, 1990; APA, Division of Evaluation, Measurement, and Statistics, 1993; Prediger, 1993; Sackett, & Wilk, 1994). В добавление к ним все более доступными становятся инструкции по корректной психологической оценке представителей разных меньшинств (см. главу 9; Dana, 1996a; Sattler, 1988, chaps. 19, & 20; Suzuki et al., 1996; Valencia, & Lopez, 1992). В докладах и отчетах, подготовленных при содействии Национального научно-исследовательского совета, Бюро технической оценки проектов (*Office of Technology Assesment*) и других подоб-

[1] Хотя женщины представляют статистическое большинство в структуре населения США, в правовом, трудовом и некоторых других отношениях они разделяют с меньшинствами многие из их проблем. Вот почему, когда термин «меньшинство» (*minority*) употребляется в этом смысле, предполагается, что он включает и женщин.

ных групп (упоминавшихся ранее в этой главе), проанализирована полемика по поводу тестов в свете современной социальной обстановки и представлены сбалансированные позиции в отношении функций тестирования.

Много внимания уделяется снижению тестовых показателей из-за возможного влияния культурных условий на развитии способностей, интересов, мотивации, аттитюдов и других психологических особенностей представителей меньшинств. Некоторые из предложенных решений этой проблемы отражают неверное понимание сущности и функции психологических тестов. Различия прошлого опыта групп или отдельных лиц неизбежно проявляются при выполнении тестов. Каждый психологический тест измеряет выборку поведения. Коль скоро культура влияет на поведение, ее влияние будет и должно обнаруживаться тестами. Если мы исключим из теста все культурные различия, мы тем самым можем уменьшить его валидность в той области поведения, для оценки которой он предназначен. В этом случае тест не сможет обеспечить нас информацией, необходимой для исправления тех условий, которые ухудшили его выполнение.

Методы тестирования для специфических популяций и их теоретическое обоснование более полно обсуждались в главах 9 и 12. Специальный анализ понятия «систематическая ошибка» был дан в главе 6, в связи с измерением валидности теста. В этой главе преимущественно рассматриваются профессиональные проблемы и социальные последствия тестирования меньшинств.

Правовое регулирование. После 1960 г. наблюдалось бурное развитие событий, имеющих отношение к тестированию меньшинств в сфере образования и трудоустройства. К этим событиям относятся законодательные меры, директивы исполнительной власти и судебные решения. Законы, касающиеся образовательного тестирования, приводились и кратко рассматривались в главах 9 и 17; обзор современных тенденций и проблем тестирования по приказу вышестоящих организаций представлен в работе Linn, & Gronlund (1995, chap. 18).[1]

В области трудоустройства суды стали играть все большую роль в толковании и применении законов о гражданских правах. Последствия нескольких знаменитых судебных прецедентов широко обсуждались в литературе по тестированию и кадровой работе психологами, юристами и лицами, имеющими психологическое и юридическое образование (см., например, APA, CPTA, 1988; Bersoff, 1983, 1984; Bruyčre, & O'Keeffe, 1994; Hollander, 1982; Merenda, 1995; Meyers, 1992; Wigdor, 1982). Имеющее самое прямое отношение к обсуждаемому вопросу федеральное законодательство обеспечивается Разделом VII Закона о гражданских правах от 1964 г. (P.L. 88–352), называемым также Законом о равных возможностях трудоустройства, вместе с последующими поправками, Законом о гражданских правах от 1991 г. (P.L. 102–166) и Законом об инвалидах-американцах от 1990 г. (P.L. 101–336). Обязанность по контролю за исполнением этих законов и право принуждения к их исполнению возлагается, главным образом, на Комиссию по вопросу равных возможностей занятости (*Equal Employment Opportunity Commission* [*EEOC*]), которая разрабатывает и распространяет для этой цели руководящие документы. В 1978 г., в интересах упрощения процедуры и улучшения координации, *EEOC*, Комиссией по государственной гражданской службе

[1] Обсуждение некоторых важных судебных решений в области психопедагогической оценки можно найти в работах Ayers, Day, & Rotatori (1990) и Reschly (1988).

(ныне — Служба управления кадрами США) и министерствами юстиции, труда и финансов были совместно приняты Единые правила проведения отбора наемных работников (*Uniform Guidelines on Employee Selection Procedures*).[1]

Закон о равных возможностях трудоустройства запрещает дискриминацию на основе таких признаков, как раса, цвет кожи, религиозные убеждения, пол или национальное происхождение, в процедурах отбора, приводящих к принятию решений о найме на работу. Эти предписания обязательны для отдельных работодателей (как частных, так и государственных), профсоюзов, бюро по трудоустройству, отделов аттестации и лицензирования. В тех случаях, когда применение теста или другой методики отбора дает существенно более высокий процент отказов в приеме на работу кандидатов из меньшинств по сравнению с процентом отказов другим кандидатам («эффект выталкивания»),[2] полезность теста или методики должна обосновываться доказательством их валидности для данного вида работы.

Если касаться истории, то требования к приемлемой валидизации теста были определены в *Стандартах тестирования*, *Принципах валидизации и использования методов отбора персонала* (*Principles for the Validation and Use of Personnel Selection Procedures* — SIOP, 1987) и других аналогичных ведомственных документах. Однако в последние два десятилетия имело место несколько прецедентов, когда чужеродные юридические соображения вторгались в психометрическую практику, особенно в связи с защитой гражданских прав. Один из этих прецедентов — юридическое соглашение, получившее известность как «Золотое правило» (см. также главу 7). Этим соглашением разрешился спор между страховой компанией «Золотое правило» и Службой тестирования в образовании (*ETS*) по поводу экзаменов, разработанных *ETS* для лицензирования страховых агентов. Соглашение предписывало, что в тех случаях, когда получается различное соотношение правильных ответов в группах меньшинства и большинства, приоритет должен отдаваться тем заданиям теста, которые обнаруживают минимальные межгрупповые различия. Хотя все это было продиктовано благими намерениями, а именно стремлением обеспечить честность испытаний и минимизировать «эффект выталкивания», соглашение «Золотое правило» вызвало горячие споры по поводу содержащихся в нем предположений о природе необъективности заданий (*item bias*) и того, в какой степени эмпирические данные оправдывают меру, предлагавшуюся в этом соглашении (APA, CPTA, 1988; Bond, 1987; Linn, & Drasgow, 1987; Rooney, 1987).

Рассматривая программу позитивных действий (*affirmative action*), «Единые правила...» указывают на то, что даже когда процедуры отбора удовлетворительно валидизированы, в случае получения непропорционально большой доли отказов для членов меньшинств следует предпринять меры для сокращения этого несоответствия до минимума. Позитивные действия подразумевают, что организация делает больше, чем просто отказывается от дискриминационной практики. Психологически, программы позитивных действий, которые в последние годы перешли в активное наступление на политической арене, можно трактовать как попытки компенсировать остаточные по-

---

[1] К настоящему времени эти «Единые правила...» устарели и явно нуждаются в критическом анализе и пересмотре. Их переработанный вариант может появиться вслед за публикацией новых *Стандартов тестирования*, ожидаемой в конце 1990-х гг. (см. главу 1).

[2] Противоречия в способах оценки «эффекта выталкивания» (*adverse impact*) в различных судебных прецедентах проанализированы Б. Лернером (B. Lerner, 1980a; см. также Ironson, Guion, & Ostrander, 1982).

следствия прошлых социальных неравенств. Применение в 1980-х гг. практики субгруппового нормирования в Батарее тестов общих способностей (*GATB*) для обеспечения сопоставимости относительного количества направлений на работу, полученных белыми, черными и испаноязычными кандидатами, несмотря на большое различие в их тестовых показателях способностей (глава 17; см. также Hartigan & Wigdor, 1989), как раз и было примером позитивных действий, нацеленных на снижение «эффекта выталкивания» теста. Эта практика, однако, вызвала настолько острую полемику, что она привела к принятию Закона о гражданских правах от 1991 г. (P.L. 102–166), запрещающего любую форму корректировки показателей на основе расы, цвета кожи, вероисповедания, пола или национального происхождения. В этой области психологического тестирования признается, что Закон от 1991 г. «имеет гораздо более серьезные последствия, чем могли представить себе члены Конгресса» (D. C. Brown, 1994, p. 927) и может серьезно ограничить применение тестов личности и физических способностей, использующих раздельные нормы для мужчин и женщин (см. также L. S. Gottfredson, 1994; Kehoe, & Tenopyr, 1994; Sackett, & Wilk, 1994). Фактически, авторы и издатели некоторых тестов уже предприняли шаги по обеспечению альтернативных способов подсчета показателей, исключающих разделение норм по полу (см., например, Gough, & Bradley, 1996).

Другая исполненная благих намерений попытка уничтожить преграды на пути к равным возможностям для всех, вызвавшая озабоченность работодателей и тех, кто заинтересован в корректной практике тестирования при приеме на работу, — Закон об инвалидах-американцах от 1990 г. (ADA, P. L. 101–336). Положения этого закона, касающиеся занятости, не позволяют работодателям до предложения работы использовать медицинские тесты и наводить справки об употреблении кандидатами наркотиков в прошлом или об их лечении от психических болезней. Разработанные *EEOC* правила и положения, касающиеся собеседований и медицинских обследований при найме на работу (1994, 1995), оставили пока открытым вопрос о том, какие из психологических и личностных тестов допустимо применять в таких ситуациях.

Противоречия между профессиональными, правовыми и этическими нормами, по-видимому, сохранятся и в будущем (см., например, D. C. Brown, 1996). И они, бесспорно, затруднят применение тестов для принятия решений в так называемых сферах «высоких интересов», к которым относятся образование и занятость. В значительной степени сложившаяся ситуация есть признак некоторого движения вперед, поскольку она подчеркивает необходимость открыто признать, что ценности вплетены в принятие любых решений, имеющих последствия, независимо от того, касаются ли они сферы научных или практических интересов. Как пишет Мессик: «Ценности изначально присущи тестированию и его результатам... Это признание делает явным то, что всегда присутствовало в скрытом состоянии, а именно: оценки валидности *являются* ценностными суждениями» (Messick, 1995, p. 748). Да, даже благонамеренные и разумные люди могут резко расходиться — и действительно расходятся — по поводу ценностей. В этот и заключается проблема.

Факторы, связанные с тестом. При тестировании лиц из разных популяций важно разделять факторы, влияющие как на сам тест, так и на критериальное поведение, и факторы, влияние которых ограничивается лишь тестом. Именно эти последние, связанные с тестом факторы (*test-related factors*), снижают его валидность. Примеры таких факторов включают опыт участия в тестах, мотивацию хорошо выполнить тест,

раппорт с тестирующим, чрезмерный акцент на скорости и любые другие переменные, влияющие на выполнение конкретного теста, но не имеющие отношения к основной области изучаемого поведения. При тестировании лиц с разным культурным происхождением или с различными дефектами необходимо сделать все возможное для ослабления действия связанных с тестом факторов (см. Sattler, 1988, chaps. 19, 20). Желательно создать сходные отношение к тесту и степень знакомства с ним, а также воспользоваться другими средствами, специально разработанными для этой цели (см. главы 1 и 9).

Специфическое содержание теста также может влиять на тестовые показатели способами, совершенно не связанными с той способностью, для оценки которой предназначен данный тест. Например, использование в тесте на арифметическое рассуждение названий или изображений предметов, неизвестных в определенной культурной среде, представляло бы связанную с тестом помеху, затрудняющую его выполнение членами такой культуры. Другой, более тонкий способ, которым специфическое содержание теста может оказывать побочное влияние на его выполнение, связан с эмоциональными реакциями и аттитюдами тестируемого. Например, рассказы или картинки, изображающие типичные для людей среднего класса семейные сцены, могут вызвать отчужденность у ребенка, живущего в необеспеченной семье. Сохранение в содержании теста половых стереотипов, наподобие изображений мужчин врачами или летчиками, а женщин — медсестрами или стюардессами, также может оказывать отрицательное воздействие. В свете этих соображений большинство издателей тестов теперь прилагают специальные усилия, чтобы очистить тест от неподходящего содержания. Фактически, проверка содержания теста на предмет возможных негативных следствий для тестируемых меньшинств является теперь общим этапом в процессе конструирования теста (см., например, EST Standards, 1981 / 1987).

Тестированию лиц с разным культурным происхождением и жизненным опытом, так же как и тестированию инвалидов, уделяется большое внимание во всех разделах *Стандартов тестирования.* Эта генеральная линия отражается в нескольких отдельных стандартах на разработку и использование тестов. В добавление к этому, специальные главы, с характерными только для них наборами стандартов, посвящены проблемам в тестировании людей с низким общественным положением и языковыми трудностями, составляющих значительную долю населения США.

Интерпретация и использование показателей теста. Безусловно, самые важные соображения, которые приходится учитывать в тестировании особых групп, да и в тестировании вообще, касаются интерпретации тестовых показателей. Наиболее частые опасения в отношении применения тестов к представителям меньшинств имеют своим источником неправильную интерпретацию показателей. Если представитель национального меньшинства получает низкий показатель по тесту способностей или отклонение в показателе по личностному тесту, важно разобраться в причинах этого. Например, низкий показатель по арифметическому тесту мог быть результатом нежелания выполнять тест, неумения хорошо читать или, среди прочих причин, недостаточного знания арифметики. Следует также обратить внимание и на тип норм, используемых при оценивании индивидуальных результатов.[1]

[1] Специальный раздел декабрьского выпуска журнала *Psychological Assessment* (December 1994) отведен под информационные и методические материалы по различным аспектам нормативной оценки.

Тесты предназначены показывать, что способен делать конкретный человек в данный момент времени. Они не могут сообщить нам, *почему* он выполняет тест именно так. Чтобы ответить на этот вопрос, нам необходимо исследовать условия его развития, мотивацию и другие релевантные обстоятельства. Тесты не могут также сообщить, на что мог бы быть способен ребенок, выросший в культурно или образовательно неблагоприятной среде, если бы он воспитывался в более благоприятной среде. К тому же тесты не могут компенсировать культурную депривацию путем исключения ее последствий из своих показателей. Напротив, тесты должны обнаруживать такие последствия, чтобы можно было предпринять соответствующие коррекционные меры. Скрывая последствия культурной депривации отказом от тестов или пытаясь изобрести тесты, нечувствительные к таким влияниям, можно только затормозить продвижение к подлинному решению социальных проблем.

Тенденция к распределению по категориям и навешиванию ярлыков, в качестве упрощенной замены понимания, все еще довольно распространена. Диагностические категории классической психиатрии, посредством которых пациенты обозначались как «параноидный шизофреник» или «маниакально-депрессивный тип», являют собой хорошо известный пример этой тенденции. Сознавая многочисленные недостатки такой системы классификации, авторы более современных руководств по психиатрической диагностике описывают расстройства различных типов и прикрепляют диагностические ярлыки к патологическим состояниям, а не к страдающим от них людям (см., например, American Psychiatric Association, 1994). Да и психологи все больше обращаются к описаниям личности. В отличие от диагностических ярлыков, эти описания сконцентрированы на происхождении и индивидуальном значении отклонений в поведении и обеспечивают более эффективную основу для терапии. Но от традиционных ярлыков удается избавиться далеко не всегда.

Еще одним примером тенденции к категоризации являются ошибки в интерпретации *IQ*. Согласно распространенному заблуждению, *IQ* служит показателем врожденного интеллектуального потенциала и представляет неизменное свойство организма. Как видно из главы 12, этот взгляд не подтверждается ни теоретическими рассуждениями, ни эмпирическими данными. Из правильно интерпретированных результатов теста интеллекта никак не следует жесткая классификация людей, напротив, интеллектуальные тесты (как и любые другие) можно сравнить с картой, на которой указано положение конкретного человека, занимаемое им в момент тестирования. В сочетании с информацией о его жизненном опыте тестовые показатели должны облегчать эффективное планирование оптимального развития индивидуума.

Объективность тестов. В ситуациях, где социальные стереотипы и предрассудки могут искажать межличностные оценки, тесты дают некоторые гарантии против фаворитизма и произвола в принятии решений. Когда движение за гражданские права набрало силу, некоторые его активные участники обратили внимание на положительную функцию, выполняемую стандартизованными тестами. Комментируя использование тестов в школах, Дж. В. Гарднер писал: «Тесты не видят, одет ли подросток в лохмотья или в твид, не слышат жаргона трущоб. Тесты выявляют интеллектуальные способности в любой из прослоек населения» (J. W. Gardner, 1961, p. 48–49).

Даже если упразднить все тесты, необходимость выбора будет по-прежнему преследовать как отдельных людей, так и целые организации. Для принятия решений пришлось бы прибегнуть к таким давно известным альтернативам, как рекомендатель-

ные письма, собеседования и средний балл. В наши дни эти альтернативные источники данных часто используют вместе с показателями тестов, но не вместо тестов. Фактически, стандартизованные тесты были внедрены в практику в качестве средства, компенсирующего ненадежность, субъективность и возможную тенденциозность этих традиционных способов. Эти альтернативы тестированию, как правило, оказывались менее точными, чем тесты, в предсказании результатов учебы или работы (Wigdor, & Garner, 1982, Pt. I, chap. 1). Более современные альтернативные способы, такие как методики оценки выполнения работы и портфельной оценки, со временем могут оказаться более эффективными по сравнению с традиционными тестами. Пока, однако, исследования с использованием этих методик свидетельствуют о том, что они не превосходят стандартизованные тесты, вместе с которыми или вместо которых они применялись для оценки представителей особых популяций, ни по валидности, ни по объективности (см. главу 17).

Огульная критика тестирования обычно не делает различий между положительным вкладом тестов в обеспечение справедливости принимаемых решений и неправильным использованием тестов в качестве упрощенных заменителей тщательно обоснованных оценок. Рассматривая тестирование в его социальном контексте, Комитет по тестированию способностей (Wigdor & Garner, 1982, Pt. I) призвал не рассматривать тесты как панацею от всех бед или, наоборот, как козла отпущения, виноватого во всех проблемах общества, и не смешивать общественные цели расширения возможностей для членов различных меньшинств со справедливостью процесса тестирования. «В поисках более справедливого общественного устройства люди поместили тестирование способностей в центр своих споров и тем самым не только прославили, но и ославили его на весь мир» (p. 239). С этим заявлением трудно не согласиться и сейчас, причем, в силу отсутствия жизнеспособных альтернатив, оно скорее всего останется правильным еще в течение долгого времени.

В общем, тесты действительно могут использоваться неправильно по отношению к меньшинствам, впрочем, как и по отношению любому другому человеку или группе. Однако когда тесты используются надлежащим образом, они выполняют важную функцию, предотвращая случайную и несправедливую дискриминацию. При оценивании социальных последствий тестирования нам необходимо тщательно оценить социальные последствия *отказа от* тестирования и вынужденной опоры на другие процедуры принятия решений, которые не столь беспристрастны ко всем, как тестирование. Кроме того, определяя последствия тестирования, мы должны быть внимательными, чтобы развести последствия правильного и неправильного использования тестов, а также отделить прямые последствия тестирования от тех, что опосредованы внешними по отношению к нему факторами (Tenopyr, 1995). В противном случае у нас есть шанс по совершенно ложным соображениям (!) отбросить за ненадобностью инструмент, который, хотя и всегда нуждался в усовершенствовании, может оказаться незаменимым.

ПРИЛОЖЕНИЕ А

Алфавитный перечень тестов и других оценочных инструментов

В этот перечень включены все обсуждаемые или упоминаемые в учебнике тесты за исключением: а) устаревших, неиспользуемых в наше время тестов, упоминавшихся в связи с историей тестирования; б) пока еще не издаваемых тестов и в) тестов, описанных в специальной литературе, получить которые можно только от их авторов. Дополнительную информацию можно найти в *Ежегодниках психических измерений* и других изданиях Института психических измерений Буроса, серии *Критические обзоры тестов (Test Critiques)* под ред. Кизера и Свитленда (Keyser, & Sweetland, 1984 – 1994), а также в других источниках, приводившихся в главе 1.

Название теста [аббревиатура] / Код издательства[1]

«Расскажи мне историю» (Tell-Me-A-Story [TEMAS]) / WPS
AAMR адаптивного поведения шкала (AAMR Adaptive Behavior Scale [ABS]) / PRO-ED
ACT-оценка (ACT Assessment) / ACT
PDI опросник службы по трудоустройству (PDI Employment Inventory) / PDI
PDI опросник службы работы с покупателями (PDI Customer Service Inventory) / PDI
TerraNova комплект (TerraNova series) / CTB
Академической оценки тест (Scholastic Assessment Test [SAT]) / ETS
Алкоголя употребления инвентарь (Alcohol Use Inventory [AUI]) / NCS
Базисный личностный опросник (Basic Personality Inventory [BPI]) / Sigma
Батарея способностей программиста ЭВМ (Computer Programmer Aptitude Battery [CPAB]) / SRA
Бейли скрининг-тест психоневрологического развития младенцев (Bayley Infant Neurodevelopmental Screener [BINS]) / TPC
Бейли шкалы развития младенцев — Вторая редакция (Bayley Scales of Infant Development — Second Edition [Bayley-II]) / TPC
Бека депрессии опросник (Beck Depression Inventory [BDI]) / TPC

[1] Приложение Б содержит полные названия и адреса издательств, выпускающих перечисленные здесь тесты, расположенные в алфавитном порядке в соответствии с кодами издательств, используемыми в этом приложении.

Бендер зрительно-моторный гештальт тест (Bender Visual Motor Gestalt Test [Bender-Gestalt]) / WPS
Беннетта тест понимания механических закономерностей (Bennett Mechanical Comprehension Test [BMCT]) / TPC
Бентона визуальной ретенции тест — Пятая редакция (Benton Visual Retention Test, Fifth Edition [BVRT] / TPC
Боэма тест базисных понятий (Boehm Test of Basic Concepts — Revised [Boehm-R]) / TPC
Брейкена базисных понятий шкала (Bracken Basic Concept Scale [BBCS]) / TPC
Британские шкалы способностей (British Ability Scales [BAS]) / NFER-Nelson
Брунинкса—Озерецкого тест двигательных умений (Bruininks-Oseretsky Test of Motor Proficiency) / AGS
Вайнлендские шкалы адаптивного поведения (Vineland Adaptive Behavior Scales [VABS]) / AGS
Вандерлика кадровый тест (Wonderlic Personnel Test) / Wonderlic
Вашингтонского университета завершения предложений тест (Washington University Sentence Completion Test [WUSCT]) / Erlbaum
Векслера индивидуальный тест достижений (Wechsler Individual Achievement Test [WIAT]) / TPC
Векслера интеллекта взрослых шкала — Пересмотренная (Wechsler Adult Intelligence Scale — Revised [WAIS-R]) / TPC
Векслера интеллекта шкала для детей — Третья редакция (Wechsler Intelligence Scale for Children — Third Edition [WISC-III]) / TPC
Векслера интеллекта шкала для дошкольников и младших школьников — Пересмотренная (Wechsler *Preschool and Primary Scale of Intelligence — Revised [WPPSI-R]*) / TPC
Вооруженных сил батарея профессиональной пригодности (Armed Services Vocational Aptitude Battery [ASVAB]) / U. S. Military
Вооруженных сил квалификационный тест (Armed Forces Qualification Test [AFQT]) / U.S. Military
Вопросник для оценки адаптации студентов к колледжу (Student Adaptation to College Questionnaire [SACQ]) WPS
Вопросник для оценки карьеры — Профессионально-техническая версия (Career Assessment Inventory — The Vocational Version [CAI-VV]) / NCS
Вопросник для оценки карьеры — Расширенная версия (Career Assessment Inventory — The Enhanced Version [CAI-EV]) / NCS
Вопросник для оценки стилей учащихся (Student Styles Questionnaire) / TPC
Встроенных фигур групповой тест (Group Embedded Figures Test) / CPP
Встроенных фигур тест (Embedded Figures Test [EFT]) / CPP
Вудкока тесты овладения чтением — Пересмотренные (Woodcock Reading Mastery Tests — Revised [WRMT-R]) / AGS
Вудкока—Джонсона психопедагогическая батарея — Пересмотренная (Woodcock-Johnson Psycho-Educational Battery — Revised [WJ-R]) / Riverside
Гаптическая шкала интеллекта (Haptic Intelligence Scale) / Stoelting
Гилфорда—Циммермана обследование темперамента (Guilford-Zimmerman Temperament Survey [GZTS]) / CPP
Гудинаф—Харриса рисования тест (Goodenough-Harris Drawing Test) / TPC
Даса—Наглиери система когнитивной оценки (Das-Naglieri Cognitive Assessment System [CAS]) / Riverside
Детской апперцепции тест (Children's Apperception Test [C.A.T.]) / CPS
Джексона личностный опросник — Пересмотренный (Jackson Personality Inventory — Revised [JPI-R]) / Sigma
Джексона профессиональных интересов обозрение (Jackson Vocational Interest Survey [JVIS]) / Sigma
Дифференциальные тесты способностей — Компьютеризованная адаптивная версия (Differential Aptitude Tests — Computerized Adaptive Edition [DAT Adaptive]) / TPC
Дифференциальные тесты способностей — Пятая редакция (Differential Aptitude Tests [DAT]) / TPC
Дифференциальные шкалы способностей (Differential Ability Scales [DAS]) / TPC
Дом-дерево-человек (House-Tree-Person [H-T-P]) / WPS
Инвентарь мнений о карьере (Career Beliefs Inventory [CBI]) / CPP
Инвентарь направлений профессиональной деятельности (Career Directions Inventory [CDI]) / Sigma
Калифорнийская колода карт для Q-сортировки (California Q-Sort Deck) / CPP

Калифорнийские диагностические тесты по математике (California Diagnostic Mathematics Tests [CDMT]) / CTB
Калифорнийские диагностические тесты чтения (California Diagnostic Reading Tests [CDRT]) / CTB
Калифорнийские тесты достижений — Пятая редакция (California Achievement Tests — Fifth Edition [CAT]) / CTB
Калифорнийский набор для Q-сортировки характеристик ребенка (California Child Q-Set) / CPP
Калифорнийский психологический опросник — Третья редакция (California Psychological Inventory — Third Edition [CPI-3]) / CPP
Карта обследования неадаптивной и адаптивной личности (Schedule for Nonadaptive and Adaptive Personality [SNAP]) / UMP
Карьеры планирования программа (Career Planning Program [CPP]) / ACT
Кауфмана краткий тест интеллекта (Kaufman Brief Intelligence Test [K-BIT]) / AGS
Кауфмана оценочная батарея для детей (Kaufman Assessment Battery for Children [K-ABC]) / AGS
Кауфмана подростков и взрослых интеллекта тест (Kaufman Adolescent and Adult Intelligence Test [KAIT]) / AGS
Кауфмана тест учебных достижений (Kaufman Test of Educational Achievement [K-TEA]) / AGS
Качества жизни опросник (Quality of Life Inventory [QOLI]) / NCS
«Ключевые элементы труда» (Work Keys) / ACT
Когнитивных способностей тест (Cognitive Abilities Test [CogAT, Form 5]) / Riverside
Колумбийская шкала умственной зрелости (Columbia Mental Maturity Scale [CMMS]) / TPC
Комплексные тесты основных навыков — Четвертая редакция (Comprehensive Tests of Basic Skills — Fourth Edition [CTBS / 4]) / CTB
Комплект для оценки понятий: Сохранение (Concept Assessment Kit — Conservation [CAK]) / EdITS
Контрольный перечень симптомов — 90 — Пересмотренный (Symptom Checklist-90 — Revised [SCL-90-R]) / NCS
Контрольный список прилагательных (Adjective Check List [ACL]) / CPP
Краткий инвентарь симптомов (Brief Symptom Inventory [BSI]) / NCS
Кросс-культурной адаптивности опросник (Cross-Cultural Adaptability Inventory [CCAI]) / NCS
Кроуфорда ловкости оперирования мелкими деталями тест (Crawford Small Parts Dexterity Test [CSPDT]) / TPC
Культурно-свободный тест интеллекта (Culture Fair Intelligence Test) / IPAT
Кьюдера общих интересов обозрение (Kuder General Interest Survey [KGIS]) / CTB
Кьюдера профессиональных интересов обозрение (Kuder Occupational Interest Survey [KOIS]) / CTB
Кьюдера профессиональных предпочтений протокол (Kuder Preference Record–Vocational [KPR-V]) / CTB
Кэмпбелла интересов и умений обозрение (Campbell Interest and Skill Survey [CISS]) / NCS
Лейтер международная шкала действия — Пересмотренная (Leiter International Performance Scale — Revised [LIPS-R]) / Stoelting
Личностный опросник для детей — Пересмотренный (Personality Inventory for Children — Revised [PIC-R]) / WPS
Личностный опросник для юношества (Personality Inventory for Youth [PIY]) / WPS
Лурия—Небраска нейропсихологическая батарея (Luria-Nebraska Neuropsychological Battery [LNNB]) / WPS
Майерс—Бриггс индикатор типов (Myers-Briggs Type Indicator [MBTI]) / CPP
Мак-Карти шкалы способностей детей (McCarthy Scales of Children's Abilities [MSCA]) / TPC
Маховер "Нарисуй человека" тест (Machover Draw-a-Person Test [D-A-P]) / Thomas
МикроКог: Оценка когнитивного функционирования (MicroCog: Assessment of Cognitive Functioning) / TPC
Миллона индекс стилей личности (Millon Index of Personality Styles [MIPS]) / TPC
Миллона клинический многоосевой опросник-III (Millon Clinical Multiaxial Inventory-III [MCMI-III]) / NCS
Миллона подростковый клинический опросник (Millon Adolescent Clinical Inventory [MACI]) / NCS
Миллона подростковый личностный опросник (Millon Adolescent Personality Inventory [MAPI]) / NCS
Миннесотский канцелярский тест (Minnesota Clerical Test [MCT]) / TPC
Миннесотский многофазный личностный опросник — 2 (Minnesota Multiphasic Personality Inventory — 2 [MMPI-2]) / UMP

Миннесотский многофазный личностный опросник — Подростковая версия (Minnesota Multiphasic Personality Inventory — Adolescent [MMPI-A]) / UMP
Многоаспектная батарея способностей (Multidimensional Aptitude Battery [MAB]) / Sigma
Нарисуй человека тест (Draw-a-Man Test) (см. Гудинаф — Харриса рисования тест)
Национальные тесты готовности — Шестая редакция (Metropolitan Readiness Tests — Sixth Edition [MRT]) / TPC
Непрерывной следящей деятельности тест (Vigil Continuous Performance Test [VIGIL]) / For Thought
Общих способностей батарея тестов (General Aptitude Test Battery [GATB]) / USES
Опросник для оценки личности (Personality Assessment Inventory [PAI]) / PAR
Опросник для оценки проявлений раздражения и раздражительности (State-Trait Anger Expression Inventory [STAXI]) / PAR
Опросник для оценки тревоги / тревожности (State-Trait Anxiety Inventory [STAI]) / CPP
Опросник для оценки тревоги / тревожности у детей (State-Trait Anxiety Inventory for Children [STAIC]) / CPP
Орлеанс—Ханна прогностический алгебраический тест (Orleans-Hanna Algebra Prognosis Test) / TPC
Отиса—Леннона школьных способностей тест — Седьмая редакция (Otis-Lennon School Ability Test — Seventh Edition [OLSAT7]) / TPC
Пересмотренный NEO-личностный опросник (Revised NEO Personality Inventory [NEO PI-R]) / PAR
Пересмотренный миннесотский бланковый тест "Доска форм" (Revised Minnesota Paper Form Board Test [RMPFBT]) / TPC
Пибоди словарный тест в картинках — Пересмотренный (Peabody Picture Vocabulary Test — Revised [PPVT-R]) / AGS
Письменные экзамены для аспирантов (Graduate Record Examinations [GRE]) / ETS
Пожилых людей апперцепции тест (Senior Apperception Test [S.A.T.]) / CPS
Портеуса лабиринты (Porteus Mazes) / TPC
Порядковые шкалы психологического развития (Ordinal Scales of Psychological Development) / UIP
Программа обследования профессиональных способностей и интересов — Вторая редакция (Occupational Aptitude Survey and Interest Schedule — Second Ed. [OASIS-2]) / PRO-ED
Программа экзаменов университетского уровня (College Level Examination Program [CLEP]) / ETS
Профессионального самоопределения опросник (Career Development Inventory [CDI]) / CPP
Профессиональных интересов инвентарь (Career Interest Inventory [CII]) / TPC
Профориентационная диалоговая система — Пересмотренная (System for Interactive Guidance Information — Revised [SIGI-PLUS]) / ETS
Равена Прогрессивные Матрицы (Raven's Progressive Matrices [RPM]) / Oxford (U. S. Distributor: TPC)
Робертса тест апперцепции для детей (Roberts Apperception Test for Children [RATC]) / WPS
Розенцвейга рисуночной фрустрации анализ (The Rosenzweig Picture-Frustration Study [P-F Study]) / PAR
Роршаха тест (Rorschach) / H & H
Роттера незаконченных предложений бланк (Rotter Incomplete Sentences Blank [RISB]) / TPC
Руководство по оценке поведения во время сеанса тестирования для WISC-III и WIAT (Guide to the Assessment of Test Session Behavior for the WISC-III and the WIAT) / TPC
Самоанализ профессиональных склонностей (Self-Directed Search [SDS]) / PAR
Система для оценки поведения детей (Behavior Assessment System for Children [BASC]) / AGS
Система оценки возрастного развития младенцев и детей раннего возраста (Infant-Toddler Developmental Assessment [IDA]) / Riverside
СИ-тест способностей к обучению (Structure of Intellect Learning Abilities Test) / WPS
Слуховой тест последовательного сложения в заданном темпе (Paced Auditory Serial Addition Test [PASAT]) / For Thought
Социального климата шкалы (Social Climate Scales) / CPP
Социальных навыков рейтинговая система (Social Skills Rating System [SSRS]) / AGS
Стронга инвентарь интересов (Strong Interest Inventory [SII]) / CPP
Стэнфорд — Бине интеллекта шкала — Четвертая редакция (Stanford — Binet Intelligence Scale — Forth Edition [SB-IV]) / Riverside
Стэнфордская программа оценки письма (Stanford Writing Assessment Program) / TPC

Стэнфордский диагностический математический тест — Третья редакция (Stanford Diagnostic Mathematics Test — Third Edition [SDMT]) / TPC
Стэнфордский диагностический тест чтения — Третья редакция (Stanford Diagnostic Reading Test — Third Edition [SDRT]) / TPC
Стэнфордский тест достижений — Восьмая редакция (Stanford Achievement Test — Eighth Edition) / TPC
Сценотест (The Scenotest) / H & H
Тематической апперцепции тест (Thematic Apperception Test [TAT]) / Harvard
Тест бригадной работы-KSA (Teamwork-KSA) / SRA
Тест достижений для учащихся американских школ — Седьмая редакция (Metropolitan Achievement Test — Seventh Edition [MAT]) / TPC
Тест когнитивных навыков — Вторая редакция (Test of Cognitive Skills — Second Edition [TCS/2]) / CTB
Тест навыков работы с видеотерминалом (CRT Skills Test) / SRA
Тест невербального интеллекта — Вторая редакция (Test of Nonverbal Intelligence — Second Edition [TONI-2]) / PRO-ED
Тест параметров внимания (Test of Variables of Attention [T.O.V.A]) / UAD
Тестовой тревожности вопросник (Test Anxiety Inventory [TAI]) / CPP
Тесты базового образования взрослых (Tests of Adult Basic Education [TABE]) / CTB
Тесты достижений и умений (Tests of Achievement and Proficiency) / Riverside
Тесты основных навыков штата Айова (Iowa Tests of Basic Skills) / Riverside
Тесты развития в обучении штата Айова (Iowa Tests of Educational Development) / Riverside
Флейшмана обзор для анализа содержания работы (Fleishman Job Analysis Survey [F-JAS]) / MRI
Форма для исследования личности (Personality Research Form [PRF]) / Sigma
Фэгана тест интеллекта младенцев (Fagan Test of Infant Intelligence) / Infantest
Халстеда—Рейтана нейропсихологических тестов батарея (Halstead-Reitan Neuropsychological Test Battery [HRB]) / RNL
Харрингтона—О'Ши система принятия карьерных решений — Пересмотренная (Harrington-O'Shea Career Decision-Making System — Revised [CDM-R]) / AGS
Хогана личностный опросник — Вторая редакция (Hogan Personality Inventory — Second Edition [HPI]) / HAS
Хольцмана чернильных пятен методика (The Holtzman Inkblot Technique [HIT]) / TPC
Ценностей шкала (The Values Scale) / CPP
Шайи—Терстоуна тест умственных способностей взрослых (Schaie-Thurstone Adult Mental Abilities Test) / CPP
Шестнадцатифакторный личностный опросник — Пятая редакция (Sixteen Personality Factor Questionnaire — Fifth Edition [16 PF]) / IPAT
Шкала Я-концепции школьника (Student Self-Concept Scale [SSCS]) / AGS
Эдвардса список личных предпочтений (Edwards Personal Preference Schedule [EPPS]) / TPC

ПРИЛОЖЕНИЕ Б

Адреса издателей, распространителей и организаций, связанных с вопросами разработки и использования тестов

AAMR	American Association on Mental Retardation 444 North Capitol Street, N.W., Suite 846 Washington, DC 20001-1512
ABPP	American Board of Professional Psychology 2100 East Broadway, Suite 313 Columbia, MO 65201-6082
ACA	American Counseling Association 5999 Stevenson Avenue Alexandria, VA 22304-3300
ACT	American College Testing Program ACT National Office 2201 North Dodge Street P.O. Box 168 Iowa City, IA 52243-0168
AERA	American Educational Research Association 1230 Seventeenth Street, N.W. Washington, DC 20036-3078
AGS	American Guidance Service, Inc. 4201 Woodland Road Circle Pines, MN 55014-1796
APA	American Psychological Association 750 First Street, N.E. Washington, DC 20002-4242
ASC	Assessment Systems Corporation 2233 University Avenue, Suite 200 St. Paul, MN 55114
ATP	Association of Test Publishers 655 Fifteenth Street, N.W, Suite 320 Washington, DC 20005
BoTA	Board on Testing and Assessment National Research Council 2101 Constitution Avenue, N.W. Washington, DC 20418
Buros	Buros Institute of Mental Measurements P.O. Box 880348 135 Bancroft Hall Lincoln, NE 68588-0348
CEEB	College Entrance Examination Board 45 Columbus Avenue New York, NY 10023-6992
CPP	Consulting Psychologists Press, Inc. 3803 East Bayshore Road P.O. Box 10096 Palo Alto, CA 94303
CPS	C.P.S, Inc. P.O. Box 83 Larchmont, NY 10538
CTB	CTB/McGraw-Hill 20 Ryan Ranch Road Monterey, CA 93940
EdITS	Educational and Industrial Testing Service P.O. Box 7234 San Diego, CA 92167
EEOC	Equal Employment Opportunity Commission 1801 L Street Washington, DC 20507
Erlbaum	Lawrence Erlbaum Associates, Inc. 10 Industrial Avenue Mahwah, NJ 07430-2262
ETS	Educational Testing Service Publications Order Services P.O. Box 6736 Princeton, NJ 08541-6736
ETS	Test Collection Mailstop 30-B Rosedale Road Princeton, NJ 08541-0001
ForThought	ForThought, Ltd. Nine Trafalgar Square Nashua, NH 03063

GRE	Graduate Record Examinations Educational Testing Service P.O. Box 6000 Princeton, NJ 08541-6000
Harcourt Brace	Harcourt Brace Educational Measurement Educational Testing Division of TPC 555 Academic Court San Antonio, TX 78204-2498
Harvard	Harvard University Press 79 Garden Street Cambridge, MA 02138
HAS	Hogan Assessment Systems, Inc. P.O. Box 521176 Tulsa, OK 74152
H & H	Hogrefe & Huber Publishers United States Office: P.O. Box 2487 Kirkland, WA 98083 Swiss Office: Verlag Hans Huber Lдnggass-Strasse 76 CH-3000 Bern 9 Switzerland
IBM	IBM K-12 Education 4111 Northside Parkway Atlanta, GA 30327
Infantest	Infantest Corporation P.O. Box 18765 Cleveland Heights, OH 44118-0765
IPAT	Institute for Personality and Ability Testing, Inc. P.O. Box 1188 Champaign, IL 61824-1188
MRI	Management Research Institute, Inc. 6701 Democracy Blvd., Suite 300 Bethesda, MD 20817
NCME	National Council on Measurement in Education 1230 Seventeenth Street, N.W. Washington, DC 20036
NCS	National Computer Systems, Inc. P.O. Box 1416 Minneapolis, MN 55440
NFER-Nelson	NFER-Nelson Publishing Company, Ltd. Darville House, 2 Oxford Road East Windsor-Berkshire, SL4 1DE United Kingdom
Oxford	Oxford Psychologists Press, Ltd. Lambourne House 311 321 Banbury Road Oxford 0X2 7JH England
PAR	Psychological Assessment Resources, Inc. P.O. Box 998 Odessa, FL 33556-0998
PDI	Personnel Decisions International 2000 Plaza VII 45 South Seventh Street Minneapolis, MN 55402-1608
PRO-ED	PRO-ED 8700 Shoal Creek Boulevard Austin, TX 78757-6897
Riverside	The Riverside Publishing Company 425 Spring Lake Drive Itasca, IL 60143
RNL	Reitan Neuropsychological Laboratory 2920 South Fourth Avenue Tucson, AZ 85713-4819
The Score	Newsletter for Division 5 of the American Psychological Association 4201 Woodland Road Circle Pines, MN 55014
Sigma	Sigma Assessment Systems, Inc. United States Office: 1110 Military Street P.O.Box 610984 Port Huron, MI 48061-0984 Canadian Office: Research Psychologists Press, Inc. 650 Waterloo Street, Suite 100 P.O. Box 3292, Station B London, ON N6A 4K3
SilverPlatter	SilverPlatter Information 100 River Ridge Drive Norwood, MA 02062-5043
SIOP	Society for Industrial and Organizational Psychology, Inc. P.O. Box 87 Bowling Green, OH 43402
SPA	Society for Personality Assessment 750 First Street, N.E. Washington, DC 20002-4242
SRA	McGraw-Hill/London House SRA Business and Industry Assessments 9701 West Higgins Road Rosemont, IL 60018-4720
Stoelting	Stoelting Company 620 Wheat Lane Wood Dale, IL 60191
Swets	Swets Test Services Heereweg 347b 2161CA Lisse Nederland
Thomas	Charles C Thomas Publisher 2600 South First Street Springfield, IL 62794-9265
TPC	The Psychological Corporation 555 Academic Court San Antonio, TX 78204-2498
UAD	Universal Attention Disorders, Inc. 4281 Katella #215 Los Alamitos, CA 90720
UIP	University of Illinois Press 1325 South Oak Street Champaign, IL 61820
UMP	University of Minnesota Press Test Division 111 Third Avenue South, Suite 290 Minneapolis, MN 55401
USES	United States Employment Service Western Assessment Research and Development Center 140 East 300 South Salt Lake City, UT 84111
U.S. Military	United States Military Entrance Processing Command Attn.: Technical Directorate 2500 Green Bay Road North Chicago, IL 60064-3094
Wonderlic	Wonderlic Personnel Test, Inc. 1509 North Milwaukee Avenue Libertyville, IL 60048-1380
WPS	Western Psychological Services 12031 Wilshire Boulevard Los Angeles, CA 90025-1251

ЛИТЕРАТУРА

ABRAHAMS, N. M., & ALF, E., JR. (1972). Pratfalls in moderator research. Journal of Applied Psychology, 56, 245–251.
ACKERMAN, P. L. (1992). Predicting individual differences in complex skills acquisition: Dynamics of ability determinants. Journal of Applied Psychology, 77, 598–614.
ACKLIN, M. W. (1995). Integrative Rorschach interpretation. Journal of Personality Assessment, 64, 235–238.
ACKLIN, M. W., MCDOWELL, C. J., & ORNDOFF, S. (1992). Statistical power and the Rorschach: 1975–1991. Journal of Personality Assessment, 59, 366–379.
ACT ASSESSMENT: USER HANDBOOK, (1995–1996). Iowa City, IA: ACT Publications.
ADAMS, R. L, PARSONS, O. A., CULBERTSON, J. L., & NIXON, S. J. (Eds.). (1996). Neuropsychology for clinical practice: Etiology, assessment, and treatment of common neurological disorders. Washington, DC: American Psychological Association.
ADCOCK, C. J. (1965). Review of Thematic Apperception Test. Sixth Mental Measurements Yearbook, 533–535.
ADLER, L. L., & GIELEN, U. P. (Eds.). (1994). Cross-cultural topics in psychology. New York: Praeger.
ADLER, N., & MATTHEWS, K. (1994). Health psychology: Why do some people get sick and some stay well? Annual Review of Psychology, 45, 229–259.
ADLER, P. A., & ADLER, P. (1994). Observational techniques. In N. K. Denzin & Y. S. Lincoln (Eds.), Handbook of qualitative research (pp. 377–392). Thousand Oaks, CA: Sage.
AGNEW, J., & MASTEN, V. L. (1994). Neuropsychological assessment of occupational neurotoxic exposure. In M. L. Bleecker &J. A. Hansen (Eds.), Occupational neurology and clinical neurotoxicology (p. 113–131). Baltimore: Williams & Wilkins.
AHLSTRÖM, K. G. (1964). Studies in spelling: I. Analysis of three different aspects of spelling ability (Rep. No. 20). Uppsala, Sweden: Uppsala University, Institute of Education.
AIKEN, L. R. (1993). Personality: Theories, research, and applications. Englewood Cliffs, NJ: Prentice Hall.
AIKEN, L. R. (1996). Assessment of intellectual functioning (2nd ed.). New York: Plenum.
AIKEN, L. S., WEST, S. G., SECHREST, L., & RENO, R. R. (1990). Graduate training in statistics, methodology, and measurement in psychology. American Psychologist, 45, 721–734.
ALBERT, R. S. (Series Ed.). (1991–1994). Creativity Research Series. Norwood, NJ: Ablex.
ALBERT, S., FOX, H. M., & KAHN, M. W. (1980). Faking psychosis on the Rorschach: Can expert judges detect malingering? Journal of Personality Assessment, 44, 115–119.
ALEXANDER, L., & JAMES, H. T. (1987). The nations report card: Improving the assessment of student achievement. Boston: Harvard Graduate School of Education, National Academy of Education.
ALLEN, R. M., & COLLINS, M. G. (1955). Suggestions for the adaptive administration of intelligence tests for those with cerebral palsy. Cerebral Palsy Review, 16, 11–14.
ALLIGER, G. M., LILIENFELD, S. O., & MITCHELL, K. E. (1996). The susceptibility of overt and covert integrity tests to coaching and faking. Psychological Science, 7, 32–39.
ALLISON, J. A. (1995). Review of the Family Environment Scale, Second Edition. Twelfth Mental Measurements Yearbook, 384–385.
ALLPORT, G. W. (1937). Personality: A psychological interpretation. New York: Holt.
ALLPORT, G. W., & ODBERT, H. S. (1936). Trait names, a psycholexical study. Psychological Monographs, 47 (1, Whole No. 211).
ALLPORT, G. W., VERNON, P. E., & LINDZEY, G. (1960). Study of Values (3rd ed.): Manual. Chicago: Riverside.
ALVARADO, N. (1994). Empirical validity of the Thematic Apperception Test. Journal of Personality Assessment, 63, 59–79.
AMELANG, M., & BORKENAU, P. (1986). The trait concept: Current theoretical considerations, empirical facts, and implications for personality inventory construction. In A. Angleitner & J. S. Wiggins (Eds.), Personality assessment via questionnaires: Current issues in theory and measurement (pp. 7–34). Berlin: Springer-Verlag.
AMERICAN ASSOCIATION ON MENTAL RETARDATION. (1992). Mental retardation; Definition, classification, and systems of supports (9th ed.). Washington, DC: Author.
AMERICAN COLLEGE TESTING PROGRAM. (1994). Counselor's manual for the ACT Career Planning Program, 3rd ed. Iowa City, IA: Author.

AMERICAN COLLEGE TESTING PROGRAM. (1995). Work Keys [Brochure]. Iowa City, IA: Author.

AMERICAN COUNSELING ASSOCIATION (ACA). (1989) Responsibilities of users of standardized tests. Alexandria, VA: Author.

AMERICAN EDUCATIONAL RESEARCH ASSOCIATION, AMERICAN PSYCHOLOGICAL ASSOCIATION, & NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION. (1985). Standards for educational and psychological testing. Washington, DC: American Psychological Association.

AMERICAN EDUCATIONAL RESEARCH ASSOCIATION, AMERICAN PSYCHOLOGICAL ASSOCIATION, & NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION (1996). Standards for educational and psychological testing. Manuscript in preparation.

AMERICAN PSYCHIATRIC ASSOCIATION. (1980). Diagnostic and statistical manual of mental disorders (3rd ed.). Washington, DC: Author.

AMERICAN PSYCHIATRIC ASSOCIATION. (1994). Diagnostic and statistical manual of mental disorders (4th ed.). Washington, DC: Author.

AMERICAN PSYCHOLOGICAL ASSOCIATION (1954). Technical recommendations for psychological tests and diagnostic techniques. Washington, DC: American Psychological Association (Also in Psychological Bulletin, 51 [2, Pt. 2].)

AMERICAN PSYCHOLOGICAL ASSOCIATION. (1982). Ethical principles in the conduct of research with human participants. Washington, DC: Author.

AMERICAN PSYCHOLOGICAL ASSOCIATION. (1987a). General guidelines for providers of psychological services. American Psychologist, 42, 712–723.

AMERICAN PSYCHOLOGICAL ASSOCIATION. (1987b). Model act for state licensure of psychologists. American Psychologist, 42, 696–703.

AMERICAN PSYCHOLOGICAL ASSOCIATION (1991). The PsycLIT Database (January, 1983—September, 1991). Washington, DC: Author.

AMERICAN PSYCHOLOGICAL ASSOCIATION. (1992). Ethical principles of psychologists and code of conduct. American Psychologist, 47, 1597–1611.

AMERICAN PSYCHOLOGICAL ASSOCIATION. (1993). Directory of the American Psychological Association. Washington, DC: Author.

AMERICAN PSYCHOLOGICAL ASSOCIATION (1994). Program: 102nd annual convention. Washington, DC: Author.

AMERICAN PSYCHOLOGICAL ASSOCIATION, AMERICAN EDUCATIONAL RESEARCH ASSOCIATION, NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION. (1974). Standards for Educational and Psychological Tests. Washington, DC: American Psychological Association.

AMERICAN PSYCHOLOGICAL ASSOCIATION, BOARD OF ETHNIC MINORITY AFFAIRS. (1990). Guidelines for providers of psychological services to ethnic, linguistic, and culturally diverse populations. Washington, DC: Author.

AMERICAN PSYCHOLOGICAL ASSOCIATION, COMMITTEE ON LEGAL ISSUES. (1996). Strategies for private practitioners coping with subpoenas or compelled testimony for client records or test data. Professional Psychology: Research and Practice, 27, 245–251.

AMERICAN PSYCHOLOGICAL ASSOCIATION, COMMITTEE ON PROFESSIONAL PRACTICE AND STANDARDS. (1993). Record keeping guidelines. American Psychologist, 48, 984–986.

AMERICAN PSYCHOLOGICAL ASSOCIATION, COMMITTEE ON PROFESSIONAL PRACTICE AND STANDARDS. (1994). Guidelines for child custody evaluations in divorce proceedings. American Psychologist, 49, 677–680.

AMERICAN PSYCHOLOGICAL ASSOCIATION, COMMITTEE ON PROFESSIONAL STANDARDS. (1981). Specialty guidelines for the delivery of services. American Psychologist, 36, 639–681.

AMERICAN PSYCHOLOGICAL ASSOCIATION, COMMITTEE ON PSYCHOLOGICAL TESTS AND ASSESSMENT. (1988). Implications for test fairness of the «Golden Rule» Company settlement. Washington, DC: Author.

AMERICAN PSYCHOLOGICAL ASSOCIATION, COMMITTEE ON PSYCHOLOGICAL TESTS AND ASSESSMENT. (1995). Statement on the use of secure psychological tests in the education of graduate and undergraduate psychology students. Washington, DC: Author.

AMERICAN PSYCHOLOGICAL ASSOCIATION, COMMITTEE ON PSYCHOLOGICAL TESTS AND ASSESSMENT. (1996). Statement on the disclosure of test data. American Psychologist, 51,644–648.

AMERICAN PSYCHOLOGICAL ASSOCIATION, DIVISION OF EVALUATION, MEASUREMENT, AND STATISTICS. (1993, January). Psychometric and assessment issues raised by the Americans with Disabilities Act (ADA). The Score Newsletter, 15, 1–2, 7–15.

AMERICAN PSYCHOLOGICAL ASSOCIATION, ETHICS COMMITTEE. (1994). Report of the Ethics Committee. American Psychologist, 49, 659–666.

AMERICAN PSYCHOLOGICAL ASSOCIATION, ETHICS COMMITTEE. (1995). Report of the Ethics Committee, 1994. American Psychologist, 50, 706–713.

AMERICAN PSYCHOLOGICAL ASSOCIATION, ETHICS COMMITTEE. (1996). Rules and procedures. American Psychologist, 51, 529–548.

AMERICAN PSYCHOLOGICAL ASSOCIATION, JOINT INTERIM COMMITTEE FOR THE IDENTIFICATION AND RECOGNITION OF SPECIALTIES AND PROFICIENCIES. (1995a). Principles for the recognition of proficiencies in psychology. Washington, DC: Author.

AMERICAN PSYCHOLOGICAL ASSOCIATION, JOINT INTERIM COMMITTEE FOR THE IDENTIFICATION AND RECOGNITION OF SPECIALTIES AND PROFICIENCIES. (1995b). Principles for the recognition of specialties in professional psychology, Washington, DC: Author.

AMERICAN PSYCHOLOGICAL ASSOCIATION, TASK FORCE ON INTELLIGENCE (1995). Intelligence: knowns and unknowns. Washington, DC: APA Science Directorate.

AMES, L. B. (1937). The sequential patterning of prone progression in the human infant. Genetic Psychology Monographs, 19, 409–460.

AMES, L. B. (1989). Arnold Gesell—Themes of his work. New York: Human Sciences Press.
ANASTASI, A. (1934). Practice and variability. Psychological Monographs, 45 (5, Whole No. 204).
ANASTASI, A. (1954). Psychological testing. New York: Macmillan.
ANASTASI, A. (1956). Age changes in adult test performance. Psychological Reports, 2, 509.
ANASTASI, A. (1958). Differential psychology (3rd ed.). New York: Macmillan.
ANASTASI, A. (Ed.). (1965). Individual differences. New York: Wiley.
ANASTASI, A. (1967). Psychology, psychologists, and psychological testing. American Psychologist, 22, 297–306.
ANASTASI, A. (1970). On the formation of psychological traits. American Psychologist 25, 899–910.
ANASTASI, A. (1971). More on heritability: Addendum to the Hebb and Jensen interchange. American Psychologist, 26, 1036–1037.
ANASTASI, A. (1972). Technical critique. In L. A. Crooks (Ed.), Proceedings of Invitational Conference on «An investigation of sources of bias in the prediction of job performance: A six-year study» (pp. 79–88). Princeton, NJ: Educational Testing Service.
ANASTASI, A. (1979). Fields of applied psychology (2nd ed.) New York: McGraw-Hill.
ANASTASI, A. (1980). Review ofR. Feuerstein et al. The dynamic assessment of retarded performers: The Learning Potential Assessment Device, theory, instruments, and techniques. Rehabilitation Literature, 41 (1-1), 28–30.
ANASTASI, A. (198 1a). Coaching, test sophistication, and developed abilities. American Psychologist, 36, 1086–1093.
ANASTASI, A. (1981b). Diverse effects of training on tests of academic intelligence. In B. F. Green (Ed.), Issues in testing: Coaching, disclosure, and ethnic bias (pp. 5–20). San Francisco: Jossey-Bass.
ANASTASI, A. (1981c). Sex differences: Historical perspectives and methodological implications. Developmental Review, 1, 187–206.
ANASTASI, A. (1983a). Evolving trait concepts. American Psychologist, 38, 175–184.
ANASTASI, A. (1983b). Traits, states, and situations: A comprehensive view. In H. Wainer & S. Messick (Eds.), Principals of modern psychological measurement: A Festschrift for Frederic M. Lord (pp. 345–356). Hillsdale, NJ: Erlbaum.
ANASTASI, A. (1983c). What do intelligence tests measure? In S. B. Anderson & J. S. Helmick (Eds.), On educational testing (pp. 5–28). San Francisco: Jossey-Bass.
ANASTASI, A. (1984a). The K-ABC in historical and contemporary perspective. Journal of Special Education, 18, 357–366.
ANASTASI, A. (1984b). Traits revisited—with some current implications. In D. P. Rogers (Ed.), Foundations of psychology: Some personal views (pp. 185–206). New York: Praeger.
ANASTASI, A. (1985a). Interpreting results from multiscore batteries. Journal of Counseling and Development, 64, 84–86.
ANASTASI, A. (1985b). Reciprocal relations between cognitive and affective development: With implications for sex differences. In T. B. Sonderegger (Ed.), Psychology and gender (Nebraska Symposium on Motivation, Vol. 32, pp. 1–35). Lincoln, NE: University of Nebraska Press.
ANASTASI, A. (1985c). Review of Kaufman Assessment Battery for Children. Ninth Mental Measurements Yearbook, Vol. 1, 769–771.
ANASTASI, A. (1985d). Some emerging trends in psychological measurement: A fifty year perspective. Applied Psychological Measurement, 9, 121–138.
ANASTASI, A. (1985e). The use of personality assessment in industry: Methodological and interpretive problems. In H. J. Bernardin & D. A. Bownas (Eds.), Personality assessment in organisations (pp. 1–20). New York: Praeger.
ANASTASI, A. (1986a). Evolving concepts of test validation. Annual Review of Psychology, 37, 1–15.
ANASTASI, A. (1986b). Experiential structuring of psychological traits. Developmental Review, 6, 181–202.
ANASTASI, A. (1986c). Intelligence as a quality of behavior. In R. J. Sternberg & D. K. Detterman (Eds.), What is intelligence? Contemporary viewpoints on its nature and definition (pp. 19–21). Norwood, NJ: Ablex.
ANASTASI, A. (1988a). Explorations in human intelligence: Some uncharted routes. Applied Measurement in Education, 1 (3), 207–213.
ANASTASI, A. (1988b). Psychological testing (6th ed.). New York: Macmillan.
ANASTASI, A. (1990a). Diversity and flexibility. The Counseling Psychologist, 18, 258–261.
ANASTASI, A. (1990b). What is test misuse? Perspectives of a measurement expert. Proceedings of the 1989 ETS Invitational Conference (pp. 15–25). Princeton, NJ: Educational Testing Service.
ANASTASI, A. (1991). The gap between experimental and psychometric orientations. Journal of the Washington Academy of Sciences, 81, 61–73.
ANASTASI, A. (1992a). Are there unifying trends in the psychologies of the 1990s? In M. E. Donnelly (Ed.), Reinterpreting the legacy of William James (pp. 29–48). Washington, DC: American Psychological Association.
ANASTASI, A. (1992b). A century of psychological science. American Psychologist, 47, 842–843.
ANASTASI, A. (1992c). Introductory remarks. In K. F. Geisinger (Ed.), Psychological testing of Hispanics (pp. 1–7). Washington, DC: American Psychological Association.
ANASTASI, A. (1993). A century of psychological testing: Origins, problems, and progress. In T. K. Fagan & G. R. VandenBos (Eds.). Exploring applied psychology: Origins and critical analysis (pp. 13–36). Washington, DC: American Psychological Association.
ANASTASI, A. (1994). Aptitude testing. Encyclopedia of human behavior (Vol. 1, pp. 211–221). San Diego, CA: Academic Press.
ANASTASI, A. (1995). Psychology evolving: Linkages, hierarchies, and dimensions. In F. Kessel (Ed.), Psychology, science and human affairs: Essays in honor of William Bevan (pp. 245–260). Boulder, CO: Westview Press.
ANASTASI, A., & DRAKE, J. (1954). An empirical comparison of certain techniques for estimating the reliability of speeded tests. Educational and Psychological Measurement, 14, 529–540.
ANDERSEN, E. B. (1983). Analyzing data using the Rasch model. In S. B. Anderson & J. S. Helmick (Eds.), On educational testing (pp. 193–223). San Francisco: Jossey-Bass.
ANDERSON, J. C., & GERBING, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. Psychological Bulletin, 103, 411–423.

ANDERSON, R. J., & SISCO, F. H. (1977). Standardization of WISC-R Performance Scale for Deaf Children. Washington, DC: Gallaudet College, Office of Demographic Studies.
ANDREW, D. M., PATERSON, D. G., & LONGSTAFF, H. P. (1979). Manual: Minnesota Clerical Test. Cleveland, OH: Psychological Corporation.
ANGLEITNER, A., JOHN, O. R, & LOHR, F. J. (1986). It's *what* you ask and *how* you ask it: An itemmetric analysis of personality questionnaires. In A. Angleitner & J. S. Wiggins (Eds.), Personality assessment via questionnaires: Current issues in theory and measurement (pp. 61–107). Berlin: Springer-Verlag.
ANGLEITNER, A., & WIGGINS, J. S. (Eds.). (1986). Personality assessment via questionnaires: Current issues in theory and measurement. Berlin: Springer-Verlag.
ANGOFF, W. H. (1962). Scales with nonmeaningful origins and units of measurement. Educational and Psychological Measurement, 22, 27–34.
ANGOFF, W. H. (1974). Criterion-referencing, norm-referencing, and the SAT. College Board Review, 92, 3–5, 21.
ANGOFF, W. H. (1984). Scales, norms, and equivalent scores. Princeton, NJ: Educational Testing Service.
ANGOFF, W. H., & COWELL, W. R. (1986). An examination of the assumption that the equating of parallel forms is population-independent. Journal of Educational Measurement, 23, 327–345.
ARCHER, R. P. (1992a). MMPI-A: Assessing adolescent psychopathology. Hillsdale, NJ: Erlbaum.
ARCHER, R. P. (1992b). Review of the Minnesota Multiphasic Personality Inventory–2. Eleventh Mental Measurements Yearbook, 558–562.
ARCHER, R. P. & KRISHNAMURTHY, R. (1994). A structural summary approach for the MMPI-A: Development and empirical correlates. Journal of Personality Assessment, 63, 554–573.
ARCHER, R. P., KRISHNAMURTHY, R., &JACOBSON, J. M. (1994). MMPI-A casebook. Odessa, EL: Psychological Assessment Resources.
ARCHER, R. P., MARUISH, M., IMHOF, E. A., & PIOTROWSKI, C. (1991). Psychological test usage with adolescent clients: 1990 survey findings. Professional Psychology: Research and Practice, 22, 247–252.
ARDILA, A., ROSSELLI, M., & PUENTE, A. E. (1994). Neuropsychological evaluation of the Spanish speaker. New York: Plenum Press.
ARKES, H. R. (1993). A practical guide to decision making [Review of the book Decision-making: Its logic and practice]. Contemporary Psychology, 38, 926–927.
Army Air Forces aviation psychology program, research reports. (1947–1948). (Nos. 1–19). Washington, DC: U.S. Government Printing Office.
ARNOLD, G. E (1951). A technique for measuring the mental ability of the cerebral palsied. Psychological Service Center Journal, 3, 171–180.
ARONOW, E., & REZNIKOFF, M. (1976). Rorschach content interpretation. Orlando, FL: Grune & Stratton.
ARONOW, E., & REZNIKOFF, M. (1983). A Rorschach introduction: Content and perceptual approaches. Orlando, FL: Grune & Stratton.
ARONOW, E., REZNIKOFF, M., & MORELAND, K. (1994). The Rorschach technique: Perceptual basics, content interpretation, and applications. Boston: Allyn & Bacon.
ARONOW, E., REZNIKOFF, M., & MORELAND, K. (1995). The Rorschach: Projective technique or psychometric test? Journal of Personality Assessment, 64, 213–228.
ARTHUR, W., JR., & DAY, D. V. (1991). Examination of the construct validity of alternative measures of field dependence/independence. Perceptual and Motor Skills, 72, 851–859.
ASSOCIATION OF TEST PUBLISHERS. (1996). Model guidelines for preemployment integrity testing (2nd ed.). Washington, DC: Author.
ATKINSON, D. R., MORTEN, G., & SUE, D. W. (1993). Counseling American minorities: A cross-cultural perspective (4th ed.). Madison, WI: Brown & Benchman/Wm. C. Brown.
ATKINSON, J. W. (Ed.). (1958). Motives in fantasy, action, and society. New York: Van Nostrand.
ATKINSON, J. W. (1974). Motivational determinants of intellective performance and cumulative achievement. In J. W. Atkinson & J. 0. Raynor (Eds.), Motivation and achievement (pp. 389–410). Washington, DC: Winston.
ATKINSON, J. W. (1981). Studying personality in the context of an advanced motivational psychology. American Psychologist, 36, 117–128.
ATKINSON, J. W., & BIRCH, G. (1978). An introduction to motivation (2nd ed.). New York: Van Nostrand.
ATKINSON, J. W., & FEATHER, N. T (Eds.). (1966). A theory of achievement motivation. New York: Wiley.
ATKINSON, J. W., O'MALLEY, P. M., & LENS, W. (1976). Motivation and ability: Interactive psychological determinants of intellective performance, educational achievement, and each other. In W. H. Sewell, R.M. Hauser, & D. L. Featherman (Eds.), Schooling and achievement in American society (pp. 29–60). New York: Academic Press.
ATKINSON, J. W., & RAYNOR, J. O. (Eds.). (1974). Motivation and achievement. Washington, DC: Winston.
ATKINSON, L., QUARRINGTON, B., ALP, I. E., & CYR, J. J. (1986). Rorschach validity: An empirical approach to the literature. Journal of Clinical Psychology, 42, 360–362.
AYERS, W., DAY, G. E, & ROTATORI, A. E (1990). Legal, judicial, and IEP parameters of testing. In A. E Rotatori, R. A. Fox, D. Sexton, &J. Miller (Eds.)., Comprehensive assessment in special education: Approaches, procedures, and concerns (pp. 124–144). Springfield, IL: Charles C Thomas.
AYLWARD, G. P. (1992). Review of Differential Ability Scales. Eleventh Mental Measurements Yearbook, 281–282.
AYLWARD, G. P. (1994). Practitioner's guide to developmental and psychological testing. New York: Plenum Press.
AYLWARD, G. P. (1995). Bayley Infant Neurodevelopmental Screener: Manual. San Antonio, TX: The Psychological Corporation.
BABAD, E. Y., & BUDOFF, M. (1974). Sensitivity and validity of learning-potential measurement in three levels of ability. Journal of Educational Psychology, 66, 439–447.
BACHELOR, P. A. (1989). Maximum likelihood confirmatory factor-analytic investigation of factors within Guilford's structure of intellect model. Journal of Applied Psychology, 74, 797–804.

BAER, J. (1993). Creativity and divergent thinking: A task-specific approach. Hillsdale, NJ: Erlbaum.
BAGNATO, S. J., & NEISWORTH, J. T. (1991). Assessment for early intervention: Best practices for professionals. New York: Guilford Press.
BAILEY, D. B., JR., & WOLERY, M. (1989). Assessing infants and preschoolers with handicaps. Columbus, OH: Merrill.
BAIRD, L. L. (1985). Field trial of a user-oriented adaptation of the inventory of documented accomplishments as a tool in graduate admissions (ETS Res. Rep. 85–13). Princeton, NJ: Educational Testing Service.
BAKER, E. L., & O'NEIL, H. E, JR. (Eds.). (1994). Technology assessment in education and training. Hillsdale, NJ: Erlbaum.
BAKER, E. L., O'NEIL, H. E, & LINN, R. L. (1993). Policy and validity prospects for performance-based assessment. American Psychologist, 48, 1210–1218.
BAKER, F. B. (1989). Computer technology in test construction and processing. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 409–428). New York: American Council on Education/Macmillan.
BAKER, R. W., & SIRYK, B. (1989). SACQ – Student Adaptation to College Questionnaire: Manual. Los Angeles: Western Psychological Services.
BALLER, W. R., CHARLES, D. C., & MILLER, E. L. (1967). Mid-life attainment of the mentally retarded: A longitudinal study. Genetic Psychology Monographs, 75, 235–329.
BALMA, M.J. (1959). The concept of synthetic validity. Personnel Psychology, 12, 395–396.
BALTES, P. B. (1968). Longitudinal and cross-sectional sequences in the study of age and generation effects. Human Development, 11, 145–171.
BALTES, P. B., CORNELIUS, S. W., SPIRO, A. III, NESSELROADE, J. R., & WILLIS, S. L. (1980). Integration vs. differentiation of fluid-crystallized intelligence in old age. Developmental Psychology, 16, 625–635.
BALTES, P. B., REESE, H. W., & LIPSITT, L. P. (1980). Life-span developmental psychology. Annual Review of Psychology, 31, 65–110.
BALZER, W. K., & SULSKY, L. M. (1992). Halo and performance appraisal research: A critical examination. Journal of Applied Psychology, 77, 975–985.
BANDURA, A. (1969). Principles of behavior modification. New York: Holt, Rinehart & Winston.
BANDURA, A. (1982). Self-efficacy mechanism in human agency. American Psychologist, 37, 720–725.
BANDURA, A. (1986). Social foundations of thought and action: A social cognitive theory. Englewood Cliffs, NJ: Prentice Hall.
BANDURA, A. (Ed.). (1995). Self-efficacy in changing societies. New York: Cambridge University Press.
BANNISTER, D. (Ed.). (1985). Issues and approaches in personal construct theory. Orlando, FL: Academic Press.
BANNISTER, D., & MAIR, J. M. M. (1968). The evaluation of personal constructs. Orlando, FL: Academic Press.
BARENDS, A., WESTEN, D., LEIGH, J., SILBERT, D., & BYERS, S. (1990). Assessing affect-tone of relationship paradigms from TAT and interview data. Psychological Assessment, 2, 329–332.
BARKLEY, R. A. (1991). The ecological validity of laboratory and analogue assessment methods of ADHD symptoms. Journal of Abnormal and Child Psychology, 19, 149–178.
BARNETT, D. W. (1983). Nondiscriminatory multifactored assessment: A sourcebook. New York: Human Sciences Press.
BARON, J. (1982). Personality and intelligence. In R. J. Sternberg (Ed.), Handbook of human intelligence (pp. 308–351). New York: Cambridge University Press.
BARRICK, M. R., & MOUNT, M. K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. Personnel Psychology, 44, 1–26.
BARRICK, M. R., & MOUNT, M. K. (1993). Autonomy as a moderator of the relationships between the Big Five personality dimensions and job performance. Journal of Applied Psychology, 78, 111–118.
BARRIOS, B. A. (1988). On the changing nature of behavioral assessment. In A. S. Bellack & M. Hersen (Eds.), Behavioral assessment: A practical handbook (3rd ed., pp. 3–41). New York: Pergamon Press.
BARRIOS, B. A. (1993). Direct observation. In T. H. Ollendick & M. Hersen (Eds.), Handbook of child and adolescent assessment (pp. 140–164). Boston: Allyn & Bacon.
BART, W. M., & AIRASIAN, P. W. (1974). Determination of the ordering among seven Piagetian tasks by an ordering-theoretic method. Journal of Educational Psychology, 66, 277–284.
BARTLETT, C. J., & EDGERTON, H. A. (1966). Stanine values for ranks for different numbers of things ranked. Educational and Psychological Measurement, 26, 287–289.
BARTRAM, D. (1993). Emerging trends in computer-assisted assessment. In H. Schuler, J. L. Farr, & M. Smith (Eds.), Personnel selection and assessment: Individual and organizational perspectives (pp. 267–288). Hillsdale, NJ: Erlbaum.
BASS, B. M. (1954). The leaderless group discussion. Psychological Bulletin, 52, 465–492.
BASS, B. M. (1990). Bass and Stodgill's handbook of leadership (3rd ed.). New York: Free Press.
BATCHELOR, E. S., J. R., & DEAN, R. S. (Eds.). (1996). Pediatric neuro-psychology: Interfacing assessment and treatment for rehabilitation. Boston: Allyn & Bacon.
BAUER, R. M. (1994). The flexible battery approach to neuropsychological assessment. In R. D. Vanderploeg (Ed.), Clinicans guide to neuropsychological assessment (pp. 259–290). Hillsdale, NJ: Erlbaum.
BAUGHMAN, E. E. (1951). Rorschach scores as a function of examiner difference. Journal of Projective Techniques, 15, 243–249.
BAUMEISTER, R. F. (Ed). (1993). Self-esteem: The puzzle of low self-regard. New York: Plenum Press.
BAUMEISTER, R. F, & TICE, D. M. (1988). Metatraits. Journal of Personality, 56, 571–598.
BAYLEY, N. (1955). On the growth of intelligence. American Psychologist, 10, 805–818.
BAYLEY, N. (1970). Development of mental abilities. In P. H. Mussen (Ed.), Carmichael's manual of child psychology (Vol. 1, pp. 1163–1209). New York: Wiley.
BAYLEY, N. (1993). Bayley Scales of Infant Development Second Edition: Manual, San Antonio, TX: Psychological Corporation.
BAYLEY, N., & ODEN, M. H. (1955). The maintenance of intellectual ability in gifted adults. Journal of Gerontology, 10, 91–107.
BAYROFF, A. G., & FUCHS, E. F. (1970). Armed Services Vocational Aptitude Battery (Tech. Res. Rep. 1161). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

BEAIL, N. (Ed.). (1985). Repertory grid technique and personal constructs: Applications in clinical and educational settings. Cambridge, MA: Brookline Books.
BECK, A. T, & STEER, R. A. (1993). Beck Depression Inventory: Manual. San Antonio, TX: Psychological Corporation.
BEDNAR, R. L., & PETERSON, S. R. (1995). Self-esteem: Paradoxes and innovations in clinical theory and practice (2nd ed.). Washington, DC: American Psychological Association.
BEILIN, H., & PUFALL, P. (Eds.). (1992). Piaget's theory: Prospects and possibilities. Hillsdale, NJ: Erlbaum.
BEJAR, I. I. (1980). Biased assessment of program impact due to psychometric artifacts. Psychological Bulletin, 87, 513–524.
BEJAR, I. I. (1985). Speculations on the future of test design. In S. E. Embretson (Ed.), Test design: Developments in psychology and psychometrics (pp. 279–294). Orlando, FL: Academic Press.
BEJAR, I. I. (1991). A generative approach to psychological and educational measurement. (Res. Rep. No. 91–20). Princeton, NJ: Educational Testing Service.
BEJAR, I. I., STABLER, E. P., & CAMP, R. (1987). Syntactic complexity and psychometric difficulty: A preliminary investigation. (Res. Rep. No. 87–25). Princeton, NJ: Educational Testing Service.
BELCHER, M. J. (1992). Review of the Wonderlic Personnel Test. Eleventh Mental Measurements Yearbook, 1044–1046.
BELL, A., & ZUBEK, J. (1960). The effect of age on the intellectual performance of mental defectives. Journal of Gerontology, 15, 285–295.
BELL, F. O, HOFF, A. L., & HOYT, K. B. (1964). Answer sheets do make a difference. Personnel Psychology, 17, 65–71.
BELLACK, A. S., & HERSEN, M. (Eds.). (1988). Behavioral assessment: A practical handbook (3rd ed.). New York: Pergamon Press.
BELLAK, L. (1992). Projective techniques in the computer age. Journal of Personality Assessment, 58, 445–453.
BELLAK, L. (1993). The Thematic Apperception Test, the Children's Apperception Test, and the Senior Apperception Test in clinical use (5th ed.). Boston: Allyn & Bacon.
BELLAK, L., & BELLAK, S. S. (1973). Manual: Senior Apperception Technique. Larchmont, NY: C. P. S.
BELLAK, L., & HURVICH, M. S. (1966). A human modification of the Children's Apperception Test (CAT-H). Journal of Projective Techniques and Personality Assessment, 30, 228–242.
BELMONT, J. M., & BUTTERFIELD, E. C. (1977). The instructional approach to developmental cognitive research. In R. V. Kail, Jr., & J. W. Hagen (Eds.), Perspectives on the development of memory and cognition (pp. 437–481). Hillsdale, NJ: Erlbaum.
BEM, D. J., & FUNDER, D. C. (1978). Predicting more of the people more of the time: Assessing the personality of situations. Psychological Review, 85, 485–501.
BENASICH, A. A., & BEJAR, I.I. (1992). The Fagan Test of Infant Intelligence: A critical review. Journal of Applied Developmental Psychology, 13, 153–171.
BENDER, L. (1938). A visual motor Gestalt test and its clinical use. American Orthopsychiatric Association, Research Monographs, No. 3.
BENES, K. M. (1995). Review of the Social Skills Rating System. Twelfth Mental Measurements Yearbook, 965–967.
BENGTSON, V. L., & SCHAIE, K. W. (Eds.). (1989). The course of later life: Research and reflections. New York: Springer.
BENNETT, G. K. (1994). Manual: BMCT-Bennett Mechanical Comprehension Test (2nd ed.). San Antonio, TX: Psychological Corporation.
BENNETT, G. K., SEASHORE, H. G., & WESMAN, A. G. (1984). Differential Aptitude Tests: Technical Supplement. San Antonio, TX: Psychological Corporation.
BENNETT, R. E. (1993). On the meanings of constructed response. In R. E. Bennett & W. C. Ward (Eds.), Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment (pp. 1–27). Hillsdale, NJ Erlbaum.
BENNETT, R. E., ROCK, D. A., & NOVATKOSKI, I. (1989). Differential item functioning on the SAT-M Braille Edition. Journal of Educational Measurement, 26, 67–79.
BENNETT, R. E., & WARD, W. C. (Eds.). (1993). Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment. Hillsdale, NJ: Erlbaum.
BEN-PORATH, Y. S., & BUTCHER, J. N. (1986). Computers in personality assessment: A brief past, an ebullient present, and an expanding future. Computers in Human Behavior, 2, 163–182.
BEN-PORATH, Y. S., & TELLEGEN, A. (1995). How (not) to evaluate the comparability of MMPI and MMPI–2 profile configurations: A reply to Humphrey and Dahlstrom. Journal of Personality Assessment, 65, 52–58.
BENTLER, P. M. (1985). Theory and implementation of EQS: A structural equations program. Los Angeles: BMDP Statistical Software.
BENTLER, P. M. (1988). Causal modeling via structural equation modeling. In J. R. Nesselroade & R. B. Cattell (Eds.), Handbook of multivariate experimental psychology (2nd ed., pp. 317–335). New York: Plenum Press.
BENTLER, P. M. (1990). Comparative fit indexes in structural models. Psychological Bulletin, 107, 238–246.
BENTON, A. L. (1994). Neuropsychological assessment. Annual Review of Psychology, 45, 1–23.
BERG, I. A. (1967). The deviation hypothesis: A broad statement of its assumptions and postulates. In I. A. Berg (Ed.), Response set in personality assessment (pp. 146–190). Chicago: Aldine.
BERK, R. A. (Ed.). (1982). Handbook of methods for detecting test bias. Baltimore: Johns Hopkins University Press.
BERK, R. A. (Ed.). (1984a). A guide to criterion-referenced test construction. Baltimore: Johns Hopkins University Press.
BERK, R. A. (1984b). Selecting the index of reliability. In R. A. Berk (Ed.), A guide to criterion-referenced test construction (pp. 231–266). Baltimore: Johns Hopkins University Press.
BERK, R. A. (1986). Minimum competency testing: Status and potential. In B. S. Plake & J. C. Witt (Eds.), The future of testing (pp. 89–144). Hillsdale, NJ: Erlbaum.
BERKAY, P. J. (1993). The adaptation of assessment center group exercises for deaf job applicants. Journal of the American Deafness and Rehabilitation Association, 27, 16–24.
BERMAN, J. J. (Ed.). (1990). Cross-cultural perspectives (Nebraska Symposium on Motivation, 1989). Lincoln: University of Nebraska Press.

BERNARDIN, H. J., & BOWNAS, D. A. (Eds.). (1985). Personality assessment in organizations. New York: Praeger.
BERNARDIN, H. J., & BUCKLEY, M. R. (1981). Strategies in rater training. Academy of Management Review, 6, 205–212.
BERNE, E. (1961). Transactional analysis in psychotherapy. New York: Grove Press.
BERNE, E. (1966). Principles of group treatment. New York: Oxford University Press.
BERNSTEIN, L. (1956). The examiner as an inhibiting factor in clinical testing. Journal of Consulting Psychology, 20, 287–290.
BERRY, D. T, WETTER, M. W., & BAER, R. A. (1995). Assessment of malingering. In J. N. Butcher (Ed.), Clinical personality assessment: Practical approaches (pp. 236–248). New York: Oxford University Press.
BERRY, J. W. (1972). Radical cultural relativism and the concept of intelligence. In L. J. Cronbach & P. J. D. Drenth (Eds.), Mental tests and cultural adaptations (pp. 77–88). The Hague: Mouton.
BERRY, J. W. (1976). Human ecology and cognitive style: Comparative studies in cultural and psychological adaptation. Beverly Hills, CA: Sage.
BERRY, J. W. (1983). Textured contexts: Systems and situations in cross-cultural psychology. In S. H. Irvine & J. W. Berry (Eds.), Human assessment and cultural factors (pp. 117–125). New York: Plenum Press.
BERRY, J. W., POORTINGA, Y. H., SEGALL, M. H., & DASEN, P. R. (1992). Cross-cultural psychology: Research and applications. New York: Cambridge University Press.
BERSOFF, D. N. (1981). Testing and the law. American Psychologist, 36, 1047–1056.
BERSOFF, D. N. (1983). Regarding psychologists testily: The legal regulation of psychological assessment. In C. J. Scheirer & B. L. Hammonds (Eds.), Psychology and law (pp. 37–88). Washington, DC: American Psychological Association.
BERSOFF, D. N. (1984). Social and legal influences on test development and usage. In B. S. Plake (Ed.), Social and technical issues in testing: Implications for test construction and usage (pp. 87–109). Hillsdale, NJ: Erlbaum.
BERSOFF, D. N. (1995). Ethical conflicts in psychology. Washington, DC: American Psychological Association.
BERSOFF, D. N., & HOFER, P. J. (1991). Legal issues in computerized psychological testing. In T. B. Gutkin & S. L. Wise (Eds.), The computer and the decision-making process (pp. 225–243). Hillsdale, NJ: Erlbaum.
BERTINI, M., PIZZAMIGLIO, L., & WAPNER, S. (1985). Field dependence in psychological theory, research, and application: Two symposia in memory of Herman A. Witkin. Hillsdale, NJ: Erlbaum.
BETZ, N. E. (1995). Gender-related individual differences variables: New concepts, methods, and measures. In D. Lubinski & R. V. Dawis (Eds.), Assessing individual differences in human behavior (pp. 119–143). Palo Alto, CA: Davies-Black.
BEUTLER, L. E., & BERREN, M. R. (1995). Integrative assessment of adult personality. New York: Guilford Press.
BIERI, J. (1971). Cognitive structures in personality. In H. M. Schroder & P Suedfeld (Eds.), Personality theory and information processing (pp. 178–208). New York: Ronald Press.
BIERI, J., ATKINS, A. L., BRIAR, S., LEAMAN, R. L., MILLER, H., & TRIPODI, T. (1966). Clinical and social judgment: The discrimination of behavioral information. New York: Wiley.
BIERMAN, K. L. (1990). Using the clinical interview to assess children's interpersonal reasoning and emotional understanding. In C. R. Reynolds & R. W. Kamphaus (Eds.), Handbook of psychological and educational assessment of children: Personality, behavior, and context (pp. 204–219). New York: Guilford Press.
BINET, A., & HENRI, V. (1895). La psychologie individuelle. Annйe Psychologique, 2, 411–463.
BINET, A., & SIMON, TH. (1905). Methodes nouvelles pour le diagnostic du niveau intellectuel des anormaux. Année Psychologique, 11, 191–244.
BIRNS, B., & GOLDEN, M. (1972). Prediction of intellectual performance at 3 years from infant test and personality measures. Merrill-Palmer Quarterly, 18, 53–58.
BIRREN, J. E., & BENGTSON, V. L. (1988). Emergent theories of aging. New York: Springer.
BIRREN, J. E., CUNNINGHAM, W. R., & YAMAMOTO, K. (1983). Psychology of adult development and aging. Annual Review of Psychology, 34, 543–575.
BIRRENJ. E., & SCHAIE, K. W. (1991). Handbook of the psychology of aging (3rd ed.). San Diego, CA: Academic Press.
BISKIN, B. H. (1992). Review of the State-Trait Anger Expression Inventory, Research Edition. Eleventh Mental Measurements Yearbook, 868–869.
BLACK, A. M., FEUER, M. J., GUIDROZ, K., & LESGOLD, A. M. (Eds.). (1996). Transitions in work and learning: Implications for assessment. Washington, DC: National Academy Press.
BLAGG, N. (1991). Can we teach intelligence? A comprehensive evaluation of Feuerstein's Instrumental Enrichment Program. Hillsdale, NJ: Erlbaum.
BLAHA, J., & WALLBROWN, E H. (1991). Hierarchical factor structure of the Wechsler Preschool and Primary Scale of Intelligence-Revised. Psychological Assessment, 3, 455–463.
BLANCHARD, W. H. (1968). The consensus Rorschach: Background and development. Journal of Projective Techniques and Personality Assessment, 32, 327–330.
BLASCOVICH, J., & TOMAKA, J. (1991). Measures of self-esteem. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), Measures of personality and social psychological attitudes. San Diego, CA: Academic Press.
BLATT, S. J. (1990). The Rorschach: A test of perception or an evaluation of representation. Journal of Personality Assessment, 55, 394–416.
BLEICHRODT, N., & DRENTH, P. J. D. (Eds.). (1991). Contemporary issues in cross-cultural psychology. Amsterdam: Swets & Zeitlinger.
BLOCK, J. (1965). The challenge of response sets: Unconfounding meaning, acquiescence, and social desirability in the MMPJ. New York: Irvington.
BLOCK, J. (1978). The Q sort method in personality assessment and psychiatric research. Palo Alto, CA: Consulting Psychologists Press. (Original work published 1961)
BLOCK, J. (1995). A contrarian view of the five-factor approach to personality description. Psychological Bulletin, 117, 187–215.
BLOOM, B. S. (1976). Human characteristics and school learning. New York: McGraw-Hill.
BLOOM, B. S., & BRODER, L. (1950). Problem-solving processes of college students. Chicago: University of Chicago Press.
BOARD ON TESTING AND ASSESSMENT. (1995). Evaluation of the U.S. Employment Service workplan for the GATB improvement project. Washington, DC: National Academy Press.

BOCHNER, S. (1986). Observational methods. In W. J. Lonner &J. W. Berry (Eds.), Field methods in cross-cultural research (pp. 165–201). Beverly Hills, CA: Sage.
BOCK, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika, 37, 29–51.
BOEHM, A. E. (1985). Review of Home Observation for Measurement of the Environment. Ninth Mental Measurements Yearbook, Vol. 1, 663–665.
BOER, E., & DUNN, J. (Eds.). (1992). Children's sibling relationships: Developmental and clinical issues. Hillsdale, NJ: Erlbaum.
BOLIG, E. E., & DAY, J. D. (1993). Dynamic assessment and giftedness: The promise of assessing training responsiveness. Roeper Review, 16, 110–113.
BOLLEN, K. A. (1989). Structural equations with latent variables. New York: Wiley.
BOLLEN, K. A., & LONG, J. S. (Eds.). (1993). Testing structural equation models. Newbury Park, CA: Sage.
BOLLER, E, & GRAFMAN, J. (Series Eds.). (1988-1995). Handbook of neuropsychology. Amsterdam: Elsevier.
BOLTON, B. (1994). [Review of the General Aptitude Test Battery]. In J. T. Kapes, M. M. Mastie, & E. A. Whitfield (Eds.), A counselor's guide to career assessment instruments (3rd ed., pp. 117–123). Alexandria, VA: National Career Development Association.
BOLTON, T. L. (1891–1892). The growth of memory in school children. American Journal of Psychology, 4, 362–380.
BOND, L. (1981). Bias in mental tests. In B. E Green (Ed.), Issues in testing: Coaching, disclosure, and ethnic bias (pp. 55–77). San Francisco: Jossey-Bass.
BOND, L. (1987). The Golden Rule settlement: A minority perspective. Educational Measurement: Issues & Practice, 6, 18–20.
BOND, L. (1989). The effects of special preparation on measures of scholastic ability. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 429–444). New York: American Council on Education/Macmillan.
BONJEAN, C. M., HILL, R. J. & MCLEMORE, S. D. (1967). Sociological measurement: An inventory of scales and indices. San Francisco: Chandler.
BONNES, M., & SECCHIAROLI, G. (1995). Environmental psychology: A psychosocial introduction. Thousand Oaks, CA: Sage.
BORGEN, R H. (1986). New approaches to the assessment of interests. In W. B. Walsh & S. H. Osipow (Eds.), Advances in vocational psychology, Vol. 1. The assessment of interests (pp. 83–125). Hillsdale, NJ: Erlbaum.
BORGEN, F. H. (1991). Megatrends and milestones in vocational behavior: A 20-year counseling psychology retrospective. Journal of Vocational Behavior, 39, 263–290.
BORGEN, F. H., & DONNAY, D. A. C. (1996). Slicing the vocational interest pie one more time: Comment on Tracey and Rounds (1996). Journal of Vocational Behavior, 48, 42–52.
BORGEN, F., & GRUTTER, J. (1995). Where do I go next? Using your Strong results to manage your career. Palo Alto, CA: Consulting Psychologists Press.
BORING, E. G. (1950). A history of experimental psychology (Rev. ed.). New York: Appleton-Century Crofts.
BORMAN, W. C. (1979). Format and training effects on rating accuracy and rating errors. Journal of Applied Psychology, 64, 410–421.
BORMAN, W. C. (1991). Job behavior, performance, and effectiveness. In M. D. Dunnette & L. M. Hough (Eds.), Handbook of industrial and organizational psychology (2nd ed., Vol. 2, pp. 271–326). Palo Alto, CA: Consulting Psychologists Press.
BORNSTEIN, M. H. (Ed). (1991). Cultural approaches to parenting. Hillsdale, NJ: Erlbaum.
BORNSTEIN, M. H., & KRASNEGOR, N. A. (Eds.). (1989). Stability and continuity in mental development: Behavioral and biological perspectives. Hillsdale, NJ: Erlbaum.
BORNSTEIN, R. E., ROSSNER, S. C., HILL, E. L., & STEPANIAN, M. L. (1994). Face validity and fakability of objective and projective measures of dependency. Journal of Personality Assessment, 63, 363–386.
BOTWINICK, J. (1984). Aging and behavior: A comprehensive integration of research findings (3rd ed.). New York: Springer.
BOUDREAU, J. W. (1991). Utility analysis for decisions in human resource management. In M. D. Dunnette & L. M. Hough (Eds.), Handbook of industrial and organizational psychology (2nd ed., Vol. 2, pp. 621–745). Palo Alto, CA: Consulting Psychologists Press.
BOWER, E. M. (1969). Early identification of emotionally handicapped children in school (2nd ed.). Springfield, IL: Charles C. Thomas.
BOWMAN, M. L. (1989). Testing individual differences in ancient China. American Psychologist, 44, 576–578.
BRACKEN, B. A. (Ed.). (1991a). The assessment of preschool children with the McCarthy Scales of Children Abilities. In B.A. Bracken (Ed.), The psychoeducational assessment of preschool children (2nd ed., pp. 53–85). Boston: Allyn & Bacon.
BRACKEN, B. A. (1991b). The psychoeducational assessment of preschool children (2nd ed.). Boston: Allyn & Bacon.
BRADEN, J. P. (1985). The structure of nonverbal intelligence in deaf and hearing subjects. American Annals of the Deaf, 130, 496–501.
BRADEN, J. P. (1994). Deafness, deprivation, and IQ. New York: Plenum Press.
BRADLEY, R. H., & BRISBY, J. A. (1993). Assessment of the home environment. In J. L. Culbertson & D. J. Willis (Eds.), Testing young children: A reference guide for developmental, psychoeducational, and psychosocial assessments (pp. 128–166). Austin, TX: PRO-ED.
BRADLEY, R. H., & CALDWELL, B. M. (1984). The HOME inventory and family demographics. Developmental Psychology, 20, 315–320.
BRADLEY-JOHNSON, S. (1994). Psychoeducational assessment of students who are visually impaired or blind: Infancy through high school (2nd ed.). Austin, TX: PRO-ED.
BRADLEY-JOHNSON, S., & EVANS, L D. (1991). Psychoeducational assessment of hearing-impaired students: Infancy through high school, Austin, TX: PRO-ED.
BRADWAY, P., THOMPSON, W., & CRAVENS, R. B. (1958). Preschool IQ's after twenty-five years. Journal of Educational Psychology, 49, 278–281.

BRAITHWAITE, V. A., & SCOTT, W. A. (1991). Values. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman, (Eds.), Measures of personality and social psychological attitudes (pp. 661–753). San Diego, CA: Academic Press.
BRANSFORD, J., SHERWOOD, R., VYE, N., & RIESER, J. (1986). Teaching thinking and problem solving: Research foundations. American Psychologist, 41, 1078–1089.
BRAUTH, S. E., HALL, W. S., & DOOLING, R. J. (Eds.). (1991). Plasticity of development, Cambridge, MA: MIT Press.
BRAY, D. W. (1982). The assessment center and the study of lives. American Psychologist, 37, 180–189.
BRAY, D. W. et al. (1991). Working with organizations and their people: A guide to human resources practice. New York: Guilford Press.
BRECKLER, S. J. (1990). Applications of covariance structure modeling in psychology: Cause for concern? Psychological Bulletin, 107, 260–273.
BRELAND, H. M. (1979). Population validity and college entrance measures (College Board Res. Monog. No. 8). New York: College Entrance Examination Board.
BRENNAN, R. L. (1984). Estimating the dependability of the scores. In R. A. Berk (Ed.), A guide to criterion-referenced test construction (pp. 292–334). Baltimore: Johns Hopkins University Press.
BRENNAN, R. L. (1994). Variance components in generalizability theory. In C. R. Reynolds (Ed.), Cognitive assessment: A multidisciplinary perspective (pp. 175–207). New York: Plenum Press.
BRIDGEMAN, B. (1974). Effects of test score feedback on immediately subsequent test performance. Journal of Educational Psychology, 66, 62–66.
BRISLIN, R. W. (1993). Understanding culture's influence on behavior. Fort Worth, TX: Harcourf Brace Jovanovich.
BRODY, N. (1992). Intelligence (2nd ed.). New York: Basic Books.
BRODZINSKY, D. M. (1982). Relationship between cognitive style and cognitive development: A 2-year longitudinal study. Developmental Psychology, 18, 617–626.
BROGDEN, H. E. (1946a). An approach to the problem of differential prediction. Psychometrika, 11, 139–154.
BROGDEN, H. E. (1946b). On the interpretation of the correlation coefficient as a measure of predictive efficiency. Journal of Educational Psychology, 37, 65–76.
BROGDEN, H. E. (1951). Increased efficiency of selection resulting from replacement of a single predictor with several differential predictors. Educational and Psychological Measurement, 11, 173–196.
BROGDEN, H. E. (1954). A simple proof of a personnel classification theorem. Psychometrika, 19, 205–208.
BRONFENBRENNER, U., & CECI, S. (1994). Nature-nurture reconceptualized in developmental perspective: A bioecological model. Psychological Review, 101, 568–586.
BROUGHTON, R. (1990). The prototype concept in personality assessment. Canadian Psychology, 31, 26–37.
BROUGHTON, R., BOYES, M. C., & MITCHELL, J. (1993). DIStance-from-the-PROtotype (DISPRO) personality assessment for children. Journal of Personality Assessment, 60, 32–47.
BROWN, A. L. (1974). The role of strategic behavior in retardate memory. In N. R. Ellis (Ed.), International Review of Research in Mental Retardation (Vol. 7, pp. 55–111). New York: Academic Press.
BROWN, A. L., & CAMPIONE, J. C. (1986). Psychological theory and the study of learning disabilities. American Psychologist, 41, 1059–1068.
BROWN, A. L., CAMPIONE, J. C., WEBBER, L. S., & McGILLY, K. (1992). Interactive learning environments: A new look at assessment and instruction. In B. R. Gifford & M. C. O'Connor (Eds.), Changing assessments: Alternative views of aptitude, achievement and instruction (pp. 121–211). Boston: Kluwer.
BROWN, C. W., & GHISELLI, E. E. (1953). Percent increase in proficiency resulting from use of selective devices. Journal of Applied Psychology, 37, 341–345.
BROWN, D., BROOKS, L., & ASSOCIATES. (1996). Career choice and development (3rd ed.). San Francisco: Jossey-Bass.
BROWN, D., & CRACE R. K. (1996). Life Values Inventory: Manual and user's guide. Chapel Hill, NC: Life Values Resources.
BROWN, D. C. (1994). Subgroup norming: Legitimate testing practice or reverse discrimination? American Psychologist, 49, 927–928.
BROWN, D. C. (1995, April). Test user qualifications. The Score Newsletter, 18, 8–9.
BROWN, D. C. (1996, January). When personality matters on the job. The Score Newsletter, 19, 4–5.
BROWN, D. T. (1989). Review of the Jackson Vocational Interest Survey. Tenth Mental Measurements Yearbook, 401–403.
BROWN, L., SHERBENOU, R. J., & JOHNSEN, S. K. (1990). Test of Nonverbal Intelligence: A language-free measure of cognitive ability (2nd ed.). Austin, TX: PRO-ED.
BROWN, S. D., & LENT, R. W. (Eds.). (1992). Handbook of counseling psychology (2nd ed.). New York: Wiley.
BRUHN, A. R. (1984). Use of early memories as a projective technique. In P. McReynolds & C. J. Chelune (Eds.), Advances in psychological assessment (Vol. 6, pp. 109–150). San Francisco: Jossey-Bass.
BRUHN, A. R. (1985). Using early memories as a projective technique — The Cognitive Perceptual method. Journal of Personality Assessment, 49, 587–597.
BRUHN, A. R. (1989). The Early Memories Procedure, Bethesda, MD: Author.
BRUHN, A. R. (1990a). Cognitive-perceptual theory and the projective use of autobiographical memory. Journal of Personality Assessment, 55, 95–114.
BRUHN, A. R. (1990b). Earliest childhood memories: Vol. 1. Theory and application to clinical practice. New York: Praeger.
BRUHN, A. R. (1992a). The Early Memories Procedure: A projective test of autobiographical memory, Part 1. Journal of Personality Assessment, 58, 1–15.
BRUHN, A. R. (1992b). The Early Memories Procedure: A projective test of autobiographical memory, Part 2. Journal of Personality Assessment, 58, 326–346.
BRUHN, A. R. (1995a). Early memories in personality assessment. In J. N. Butcher (Ed.), Clinical personality assessment: Practical approaches (pp. 278–301). New York: Oxford University Press.
BRUHN, A. R. (1995b). Ideographic aspects of injury memories: Applying contextual theory to the Comprehensive Early Memories Scoring System—Revised. Journal of Personality Assessment, 65, 195–236.

BRUHN, A. R., & LAST, J. (1982). Earliest childhood memories: Four theoretical perspectives. Journal of Personality Assessment, 46, 119–127.
BRUININKS, R. H. (1978). Bruininks-Oseretsky Test of Motor Proficiency: Examiner's manual. Circle Pines, MN: American Guidance Service.
BRUYĈRE, S. M., & O'KEEFE, J. (Eds.). (1994). Implications of the Americans with Disabilities Act for psychology. Washington, DC: American Psychological Association.
BUCHWALD, A. M. (1965). Values and the use of tests. Journal of Consulting Psychology, 29, 49–54.
BUCK, J. N. (1948). The H-T-P technique, a qualitative and quantitative method. Journal of Clinical Psychology, 4, 317–396.
BUCK, J. N. (1992). House-Tree-Person protective drawing technique (H-T-P): Manual and interpretative guide (Revised by W. L. Warren). Los Angeles, CA: Western Psychological Services.
BUDOFF, M. (1987). A learning potential assessment battery. In C. S. Lidz (Ed.), Dynamic assessment: An interactive approach to evaluating learning potential (pp. 167–193). New York: Guilford Press.
BUDOFF, M., & CORMAN, L. (1974). Demographic and psychometric factors related to improved performance on the Kohs learning potential procedure. American Journal of Mental Deficiency, 78, 578–585.
BURGEMEISTER, B. B., BLUM, L. H., & LORGE, I. (1972). Columbia Mental Maturity Scale: Guide for administering and interpreting (3rd ed.). New York: Harcourt Brace Jovanovich.
BURGER, J. M. (1993). Personality (3rd ed.). Pacific Grove, CA: Brooks/Cole.
BURISCH, M. (1986). Methods of personality inventory development — A comparative analysis. In A. Angleitner & J. S. Wiggins (Eds.), Personality assessment via questionnaires: Current issues in theory and measurement (pp. 109–120). Berlin: Springer-Verlag.
BURKE, M. J. (1993). Computerized psychological testing: Impacts on measuring predictor constructs and future job behavior. In N. Schmitt, W. C. Botman et al. (Eds.). Personnel selection in organizations (pp. 203–239). San Francisco: Jossey-Bass.
BURKE, M. J., & FREDERICK, J. T. (1984). Two modified procedures for estimating standard deviations in utility analyses. Journal of Applied Psychology, 69, 482–489.
BURNHAM, P. S. (1965). Prediction and performance. In From high school to college: Readings for counselors (pp. 65–71). New York: College Entrance Examination Board.
BURNS, R. B. (1966). Age and mental ability: Retesting with thirty-three years' interval. British Journal of Educational Psychology, 36, 116.
BURNS, R. B. (1980). Relation of aptitudes to learning at different points in time during instruction. Journal of Educational Psychology, 72, 785–795.
BURNS, R. C. (1982). Self-growth in families: Kinetic Family Drawings (K-F-D) research and applications. New York: Brunner/Mazel.
BURNS, R. C., & KAUFMAN, S. H. (1970). Kinetic Family Drawings (K-F-D): An introduction to understanding children through kinetic drawings. New York: Brunner/Mazel.
BURNS, R. C., & KAUFMAN, S. H. (1972). Actions, styles, and symbols in Kinetic Family Drawings (K-F-D): An interpretative manual. New York: Brunner/Mazel.
BUROS, O. (Ed.). (1974). Tests in print II. Lincoln, NE: Buros Institute of Mental Measurements.
BUROS, O. K. (Ed.). (1975). Vocational tests and reviews. Highland Park, NJ: Gryphon Press.
BURR, V., & BUTT, T (1992). Invitation to personal construct psychology. London: Whurr.
BURT, C. (1941). The factors of the mind: An introduction to factor-analysis in psychology. New York: Macmillan.
BURT, C. (1944). Mental abilities and mental factors. British Journal of Educational Psychology, 14, 85–89.
BURT, C. (1949). The structure of the mind; a review of the results of factor analysis. British Journal of Educational Psychology, 19, 110–111; 176–199.
BURTON, R. V. (1963). Generality of honesty reconsidered. Psychological Review, 70, 481–499.
BURTT, H. E. (1931). Legal psychology. Englewood Cliffs, NJ: Prentice Hall.
BUSHE, G. R., & GIBBS, B. W. (1990). Predicting organization development consulting competence from the Myers-Briggs Type Indicator and stage of ego development. Journal of Applied Behavioral Science, 26, 337–357.
BUSS, A. R. (1973). An extension of developmental models that separate ontogenetic changes and cohort differences. Psychological Bulletin, 80, 466–479.
BUTCHER, J. N. (Ed.) (1985). Perspectives on computerized psychological assessment [Special issue]. Journal of Consulting and Clinical Psychology, 53(6).
BUTCHER, J. N. (Ed.). (1987). Computerized psychological assessment: A practitioner's guide. New York: Basic Books.
BUTCHER, J. N. (1990). MMPI–2 in psychological treatment. New York: Oxford University Press.
BUTCHER, J. N. (Ed.). (1995). Clinical personality assessment: Practical approaches. New York: Oxford University Press.
BUTCHER, J. N. (Ed.). (1996). International adaptations of the MMPI–2: Research and clinical applications. Minneapolis: University of Minnesota Press.
BUTCHER, J. N., DAHLSTROM, W. G., GRAHAM, J. R., TELLEGEN, A., & KAEMMER, B. (1989). Minnesota Multiphasic Personality Inventory–2 (MMPI–2): Manual for administration and scoring. Minneapolis: University of Minnesota Press.
BUTCHER, J. N., GRAHAM, J. R., & BEN-PORATH, Y. S. (1995). Methodological problems and issues in MMPI, MMPI–2, and MMPI-A Research. Psychological Assessment, 7, 320–329.
BUTCHER, J. N., GRAHAM, J. R., WILLIAMS, C. L., & BEN-PORATH, Y. S. (1990). Development and use of the MMPI–2 content scales. Minneapolis: University of Minnesota Press.
BUTCHER, J. N., & ROUSE, S. V. (1996). Personality: Individual differences and clinical assessment. Annual Review of Psychology, 47, 87–111.
BUTCHER, J. N., & WILLIAMS, C. L. (1992). Essentials of MMPI–2 and MMPI-A interpretation. Minneapolis: University of Minnesota Press.
BUTCHER, J. N.. WILLIAMS, C. L., GRAHAM, J. R., ARCHER, R. P., TELLEGEN, A., BEN-PORATH, Y. S., & KAEMMER, B. (1992). Minnesota Multiphasic Personality Inventory-Adolescent (MMPI-A): Manual for administration, scoring, and interpretation. Minneapolis: University of Minnesota Press.

BUTTERFIELD, E. C., NIELSEN, D., TANGEN, K. L., & RICHARDSON, M. B. (1985). Theoretically based psychometric measures of inductive reasoning. In S. E. Embretson (Ed.), Test design: Developments in psychology and psychometrics (pp. 77–147). Orlando, FL: Academic Press.
BUTTERS, N., DELIS, D. C., & LUCAS, J. A. (1995). Clinical assessment of memory disorders in amnesia and dementia. Annual Review of Psychology, 46, 493–523.
BUTTERWORTH, G. E., HARRIS, P. L., LESLIE, A. M., & WELLMAN, H. M. (Eds.). (1991). Perspectives on the child's theory of mind. Oxford, England: British Psychological Society and Oxford University Press.
BYRNE, B. M. (1996). Measuring self-concept across the life span: Issues and instrumentation. Washington, DC: American Psychological Association.
CALDWELL, B. M., & BRADLEY, R. H. (1978). Home Observation for Measurement of the Environment, Little Rock, AR: Authors.
CALDWELL, B. M., & BRADLEY, R. H. (1984). Home Observation/or Measurement of the Environment. Little Rock: University of Arkansas.
CALDWELL, O. W., & COURTIS, S. A. (1923). Then and now in education, 1845–1923. Yonkers, NY: World Book.
CAMARA, W., FREEMAN, J., & EVERSON, H. (1996). Using the SAT: Technical supplement. New York: College Entrance Examination Board. Manuscript in preparation.
CAMARA, W. J., & SCHNEIDER, D. L. (1994). Integrity tests: Facts and unresolved issues. American Psychologist, 49, 112–119.
CAMARA, W. J., & SCHNEIDER, D. L. (1995). Questions of construct breadth and openness of research in integrity testing. American Psychologist, 50, 459–460.
CAMILLI, G., & SHEPARD, L. A. (1994). Methods for identifying biased test items. Newbury Park, CA: Sage.
CAMP, R. (1993). The place of portfolios in our changing views of writing assessment. In R. E. Bennett & W. C. Ward (Eds.), Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, portfolio assessment (pp. 183–212). Hillsdale, NJ: Erlbaum.
CAMPBELL, D. R (1965). A cross-sectional and longitudinal study of scholastic abilities over twenty-five years. Journal of Counseling Psychology, 12, 55–61.
CAMPBELL, D. P. (1971). Handbook for the Strong Vocational Interest Blank. Stanford, CA: Stanford University Press.
CAMPBELL, D. P. (1974). Manual for the Strong-Campbell Interest Inventory. Stanford, CA: Stanford University Press.
CAMPBELL, D. P. (1977). Manual for the Strong-Campbell Interest Inventory (rev. ed.). Stanford, CA: Stanford University Press.
CAMPBELL, D. P., & HANSEN, J. C. (1981). Manual for the SVIB-SCII (3rd ed.). Stanford, CA: Stanford University Press.
CAMPBELL, D. P., HYNE, S. A., & NILSEN, D. L. (1992). Manual for the Campbell Interest and Skill Survey (CISS). Minneapolis, MN: National Computer Systems.
CAMPBELL, D. T. (1950). The indirect assessment of social attitudes. Psychological Bulletin, 47,15–38.
CAMPBELL, D. T. (1960). Recommendations for APA test standards regarding construct, trait, and discriminant validity. American Psychologist, 15, 546–553.
CAMPBELL, D. T, & FISKE, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 56, 81–105.
CAMPBELL, D. T, & STANLEY, J. C. (1966). Experimental and quasi-experimental designs for research, Chicago: Rand McNally.
CAMPBELL, R A., & RAMEY, C. T. (1990). The relationship between Piagetian cognitive development, mental test performance, and academic achievement in high risk students with and without early educational experience. Intelligence, 14, 293–308.
CAMPBELL, I. A. (1985). Review of the Vineland Adaptive Behavior Scales. Ninth Mental Measurements Yearbook, Vol. 2, 1660–1662.
CAMPBELL, J. P. (1990a). Modeling the performance prediction problem in industrial and organizational psychology. In M. D. Dunnette & L. M. Hough (Eds.), Handbook of industrial and organizational psychology (2nd ed., Vol. 1, pp. 687–732). Palo Alto, CA: Consulting Psychologists Press.
CAMPBELL, J. P. (1990b). An overview of the Army Selection and Classification Project (Project A). Personnel Psychology, 43, 231–239.
CAMPBELL, J. R (1994). Alternative models of job performance and their implications for selection and classification. In M. G. Rumsey, C. B. Walker, & J. H. Harris (Eds.), Personnel selection and classification (pp. 33–51). Hillsdale, NJ: Erlbaum.
CAMPBELL, J. P., CAMPBELL, R. J., & ASSOCIATES (1988). Productivity in organizations: New perspectives from industrial and organizational psychology. San Francisco: Jossey-Bass.
CAMPBELL, J. P., McCLOY, R. A., OPPLER, S. H., & SAGER, C. E. (1993). A theory of performance. In N. Schmitt, W. C. Borman, et al. (Eds.), Personnel selection in organizations (pp. 35–70). San Francisco: Jossey-Bass.
CAMPBELL, J. P., MCHENRY, J. J., & WISE, L. L. (1990). Modeling job performance in a population of jobs. Personnel Psychology, 43, 313–333.
CAMPBELL, J. T, CROOKS, L. A., MAHONEY, M. H., & ROCK, D. A. (1973). An investigation of sources of bias in the prediction of job performance: A six-year study. Princeton, NJ: Educational Testing Service.
CAMPION, M. A. (1994). Job analysis for the future. In M. G. Rumsey, C. B. Walker, & J.-H. Harris (Eds.), Personnel selection and classification (pp. 1–12). Hillsdale, NJ: Erlbaum.
CAMPIONE, J. C., & BROWN, A. L. (1979). Toward a theory of intelligence: Contributions from research with retarded children. In R. J. Sternberg & D. K. Detterman (Eds.), Human intelligence: Perspectives on its theory and measurement (pp. 139–163). Norwood, NJ: Ablex.
CAMPIONE, J. C., & BROWN, A. L. (1987). Linking dynamic assessment with school achievement. In C. S. Lidz (Ed.), Dynamic assessment: An interactive approach to evaluating learning potential (pp. 76–109). New York: Guilford Press.
CANCELLI, A. A., & ARENA, S. T. (1996). Multicultural implications of performance-based assessment. In L. A. Suzuki, P. J. Meller, & J. G. Ponterotto (Eds.), Handbook of multicultural assessment: Clinical, psychological, and educational applications (pp. 319–347). San Francisco: Jossey-Bass.

CANFIELD, A. A. (1951). The «sten» scale—A modified C-scale. Educational and Psychological Measurement, 11, 295–297.
CANTER, A. (1996). The Bender-Gestalt Test (BGT). In C. S. Newmark (Ed.), Major psychological assessment instruments (2nd ed., pp. 400–432). Boston: Allyn & Bacon.
CANTER, M. B., BENNETT, B. E., JONES, S. E., & NAGY, T. R (1994). Ethics for psychologists: A commentary on the APA ethics code. Washington, DC: American Psychological Association.
CAPITANI, E., SALA, S. D., & MARCHITTI, C. (1994). Is there a cognitive impairment in MND? A survey with longitudinal data. Schweizer Archiv fur Neurologie und Psychiatric, 145, 11–13.
CARLSON, R. (1992). Shrinking personality: One cheer for the Big Five [Review of R. R. McCrae and P. T. Costa, Jr., Personality in adulthood]. Contemporary Psychology, 37, 644–645.
CARROLL, J. B. (1963). A model of school learning. Teachers College Record, 64, 723–733.
CARROLL, J. B. (1966). Factors of verbal achievement. In A. Anastasi (Ed.), Testing problems in perspective (pp. 406–413). Washington, DC: American Council on Education.
CARROLL, J. B. (1970). Problems of measurement related to the concept of learning for mastery. Educational Horizons, 48, 71–80.
CARROLL, J. B. (1972). Stalking the wayward factors [Review of The analysis of intelligence by J. R Guilford & R. Hoepfner]. Contemporary Psychology, 17, 321–324.
CARROLL, J. B. (1987). New perspectives in the analysis of abilities. In R. R. Ronning, J. A. Glover, J. C. Conoley, & J. C. Witt (Eds.), The influence of cognitive psychology on testing (pp. 267–284). Hillsdale, NJ: Erlbaum.
CARROLL, J. B. (1992). Cognitive abilities: The state of the art. Psychological Science, 3, 266–270.
CARROLL, J. B. (1993). Human cognitive abilities: A survey of factor-analytic studies. New York: Cambridge University Press.
CARSON, K. P., & GILLIARD, D. J. (1993). Construct validity of the Miner Sentence Completion Scale. Journal of Occupational and Organizational Psychology, 66, 171–175.
CARVER, R. P. (1993). The case against statistical significance testing, revisited. Journal of Experimental Education, 61, 287–292.
CASCIO, W. R, & MORRIS, J. R. (1990). A critical analysis of Hunter, Schmidt, and Coggins (1988) «Problems and pitfalls in using capital budgeting and financial accounting techniques in assessing the utility of personnel programs.» Journal of Applied Psychology, 75, 410–417,
CASHEN, V. M., & RAMSEYER, G. C. (1969). The use of separate answer sheets by primary age children. Journal of Educational Measurement, 6, 155–158.
CASPI, A., BLOCK, J., BLOCK, J. H., KLOPP, B., LYNAM, D., MOFFITT, T. E., & STOUTHAMER-LOEBER, M. (1992). A «common language» version of the California Child Q-Set for personality assessment. Psychological Assessment, 4, 512–523.
CATTELL, R. B. (1979). Personality and learning theory: Vol. I. The structure of personality and its environment. New York: Springer.
CATTELL, R. B., CATTELL, A. K., & CATTELL, H. E. (1993). Sixteen Personality Factor Questionnaire, Fifth Edition. Champaign, IL: Institute for Personality and Ability Testing.
CAUDILL, O.B., JR., & POPE, K. S. (1995). Law and mental health professionals: California. Washington, DC: American Psychological Association.
CEGALIS, J. A., & BIRDSALL, W. (1995). Paced Auditory Serial Attention Task. Nashua, NH: ForThought.
CEGALIS, J. A., CEGALIS, S., & BOWLIN, J. (1993). Vigil/W: Continuous Performance Test. Nashua, NH: ForThought.
CHAPMAN, L. J. (1967). Illusory correlation in observational report. Journal of Verbal Learning and Verbal Behavior, 6, 151–155.
CHAPMAN, L. J., & CHAPMAN, J. P. (1967). Genesis of popular but erroneous psychodiagnostic observations. Journal of Abnormal Psychology, 72, 193–204.
CHAPMAN, L. J., & CHAPMAN, J. P. (1969). Illusory correlation as an obstacle to the use of valid psychodiagnostic signs. Journal of Abnormal Psychology, 74, 271–280.
CHARLES, D. C. (1953). Ability and accomplishment of persons earlier judged mental deficient. Genetic Psychology Monographs, 47, 3–71.
CHARLES, D. C., & JAMES, S. T. (1964). Stability of average intelligence. Journal of Genetic Psychology, 105, 105–111.
CHI, M. T. H., GLASER, R., & FARR, M. J. (Eds.). (1988). The nature of expertise. Hillsdale, NJ: Erlbaum.
CHOCA, J. P., SHANLEY, L. A., & VAN DENBURG, E. (1992). Interpretive guide to the Millon Clinical Multiaxial Inventory (MCMI). Washington, DC: American Psychological Association.
CHOJNACKI, J. T, & WALSH, W. B. (1992). The consistency of scores and configural patterns between the MMPI and MMPI–2. Journal of Personality Assessment, 59, 276–289.
CHRISTAL, R. E. (1958). Factor analytic study of visual memory. Psychological Monographs, 72 (13, Whole No. 466).
CHRISTENSEN, A. L. (1975). Luria's neuropsychological investigation. New York: Spectrum.
CLARK, K. E. (1961). Vocational interests of non-professional men. Minneapolis: University of Minnesota Press.
CLARK, K. E., & CLARK, M. B. (Eds.). (1990). Measures of leadership. West Orange, NJ: Leadership Library of America.
CLARK, L. A., McEWEN, J. L., COLLARD, L. M., & HICKOK, L. G. (1993). Symptoms and traits of personality disorder: Two new methods for their assessment. Psychological Assessment, 5, 81–91.
CLEARY, T. A. (1968). Test bias: Prediction of grades of Negro and white students in integrated colleges. Journal of Educational Measurement, 5, 115–124.
CLEARY, T A., LINN, R. L., & ROCK, D. A. (1968). An exploratory study of programmed tests. Educational and Psychological Measurement, 28, 347–349.
CLEMANS, W. V. (1958). An index of item-criterion relationship. Educational and Psychological Measurement, 18, 167–172.
COATES, S. (1972). Preschool Embedded Figures Test. Palo Alto, CA: Consulting Psychologists Press.
COFFMAN, W. E. (1985). Review of Kaufman Assessment Battery for Children. Ninth Mental Measurements Yearbook, Vol. 1, 771–773.
COGLISER, C. C., & SCHRIESHEIM, C. A. (1994). Development and application of a new approach to testing the bipolarity of semantic differential items. Educational and Psychological Measurement, 54, 594–605.

COHEN, J. (1994). The earth is round (p < 0,05). American Psychologist, 49, 997–1003.
COHEN, R. A. (1969). Conceptual styles, culture conflict, and nonverbal tests. American Anthropologist, 71, 828–856.
COHN, L. D. (1991). Sex differences in the course of personality development: A meta-analysis. Psychological Bulletin, 109, 252–266.
COLBERG, M. (1985). Logic-based measurement of verbal reasoning: A key to increased validity and economy. Personnel Psychology, 38, 347–359.
COLBERG, M., NESTER, M. A., & TRATTNER, M. H. (1985). Convergence of the inductive and deductive models in the measurement of reasoning abilities. Journal of Applied Psychology, 70, 681–694.
COLE, D. A., MAXWELL, S. E., ARVEY, R., & SALAS, E. (1993). Multivariate group comparisons of variable systems: MANOVA and structural equation modeling. Psychological Bulletin, 114, 174–184.
COLE, D. A., & WHITE, K. (1993). Structure of peer impressions of children's competence: Validation of the Peer Nomination of Multiple Competencies. Psychological Assessment, 5, 449–456.
COLE, M., & BRUNER, J. S. (1971). Cultural differences and inferences about psychological processes. American Psychologist, 26, 867–876.
COLE, N. S., & MOSS, P. A. (1989). Bias in test use. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 201–219). New York: American Council on Education/ Macmillan.
COLEMAN, J. L. (1987). Police assessment testing: An assessment center handbook for law enforcement personnel, Springfield, IL: Charles C Thomas.
COLEMAN, W., & CURETON, E. E. (1954). Intelligence and achievement: «The jangle fallacy» again. Educational and Psychological Measurement, 14, 347–351.
COLLEGE BOARD. (1995a). Admission officers handbook for the SAT Program. New York: College Entrance Examination Board.
COLLEGE BOARD. (1995b). Counselors handbook for the SAT Program. New York: College Entrance Examination Board.
COLLIGAN, R. C., OSBORNE, D., SWENSON, W. M., & OFFORD, K. P. (1983). The MMPI: A contemporary normative study. New York: Praeger.
COLLIGAN, R. C., OSBORNE, D., SWENSON, W. M., & OFFORD, K. P. (1989). The MMPI: A contemporary normative study of adults (2nd ed.). Odessa, FL: Psychological Assessment Resources.
COLLINS, B. E. (1974). Four components of the Rotter Internal-External Scale: Belief in a difficult world, a just world, a predictable world, and a politically responsive world. Journal of Personality and Social Psychology, 29, 381–391.
COLLINS, C., & MANGIERI, J. N. (Eds.). (1992). Teaching thinking: An agenda for the 21st century. Hillsdale, NJ: Erlbaum.
COLLINS, L. M., & HORN, J. L. (Eds). (1991). Best methods for the analysis of change: Recent advances, unanswered questions, future directions. Washington, DC: American Psychological Association.
COLLINS, R. C. (1993). Head Start: Steps toward a two-generation program strategy. Young Children, 48 (2), 25–73.
COLOMBO, J. (1993). Jnfant cognition: Predicting later intellectual functioning. Newbury Park, CA, Sage.
COMREY, A. L., & LEE, H. B. (1992). A first course in factor analysis (2nd ed.). Hillsdale, NJ: Erlbaum.
CONGER, A. J., & JACKSON, D. N. (1972). Suppressor variables, prediction, and the interpretation of psychological relationships. Educational and Psychological Measurement, 32, 579–599.
CONN, S. R., & RIEKE, M. L. (Eds.). (1994). The 16 PF Fifth Edition technical manual. Champaign, IL: Institute for Personality and Ability Testing.
CONNELL, J. P. (1985). A new multidimensional measure of children's perceptions of control. Child Development, 56, 1018–1041.
CONNOR, M. (1994). Training the counselor: An integrative model. London: Routledge.
CONOLEY, J. C., & WERTH, E. B.(Eds.). (1995). Family assessment. Lincoln, NE: Buros Institute of Mental Measurements.
CONSORTIUM FOR LONGITUDINAL STUDIES. (1983). As the twig is bent...: Lasting effects of preschool programs. Hillsdale, NJ: Erlbaum.
COOK, T. D., & CAMPBELL, D. T (1976). The design and conduct of quasi-experiments and true experiments in field settings. In M. D. Dunnette (Ed.), Handbook of industrial and organizational psychology (pp. 223–326). Chicago: Rand-McNally.
COOK, T. D., COOPER, H., CORDRAY, D. S., HARTMAN, H., HEDGES, L. V, LIGHT, R. J., LOUIS, T. A., & MOSTELLER, F. (1992). Meta-analysis for explanation: A casebook. New York: Russell Sage Foundation.
COOLEY, W. W., & GLASER, R. (1969). The computer and individualized instruction. Science, 166, 574–582.
COOLEY, W. W., & LOHNES, P. (1976). Evaluation research in education. New York: Wiley.
COOPER, H., & HEDGES, L. V. (Eds.). (1994). The handbook of research synthesis. New York: Russell Sage Foundation.
COSDEN, M. (1992). Review of the Draw A Person: A quantitative scoring system. Eleventh Mental Measurements Yearbook, 287–289.
COSTA, P. T., JR., & McCRAE, R. R. (1988). From catalogue to classification: Murray's needs and the five-factor model. Journal of Personality and Social Psychology, 55, 258–265.
COSTA, P. T., JR., & McCRAE, R. R. (1992a). Normal personality assessment in clinical practice: The NEO Personality inventory. Psychological Assessment, 4, 5–13.
COSTA, P. T., JR, & McCRAE, R. R. (19920). Revised NEO Personality Inventory (NEO PI-R) and NEO Fife-Factor Inventory (NEO-FFI) professional manual, Odessa, FL: Psychological Assessment Resources.
COSTA, P. T., JR., & McCRAE, R. R. (1994). Bibliography for the Revised NEO Personality Inventory and NEO Five-Factor Inventory (NEO-FFI). Odessa, FL: Psychological Assessment Resources.
COSTA, P. T., JR., & McCRAE, R. R. (1995). Domains and Facets: Hierarchical personality assessment using the Revised NEO Personality Inventory. Journal of Personality Assessment, 64, 21–50.
COSTA, P. T., JR., McCRAE, R. R., & HOLLAND, J, L. (1984). Personality and vocational interests in an adult sample. Journal of Applied Psychology, 69, 390–400.
COSTA, P. T., JR., & WIDIGER, T. A. (Eds.). (1994). Personality disorders and the Five-Factor Model of personality. Washington, DC: American Psychological Association.
COSTANTINO, G., MALGADY, R. G., & ROGLER, L. H. (1988). TEMAS (Tell-Me-A-Story): Manual. Los Angeles, CA: Western Psychological Services.

COUCH, A., & KENISTON, K. (1960). Yeasayers and naysayers: Agreeing response set as a personality variable. Journal of Abnormal and Social Psychology, 60, 151–174.
COULTON, G. E., & FEILD, H. S. (1995). Using assessment centers in selecting entry-level police officers: Extravagance or justified expense? Public Personnel Management, 24, 223–254.
COURT, J. H., & RAVEN, J. (1995). Manual for Raven's Progressive Matrices and vocabulary scales: Sect. 7. Research and references. Oxford, England: Oxford Psychologists Press.
COURTS, P. L., & McINERNEY, K. H. (1993). Assessment in higher education: Politics, pedagogy, and portfolios. Westport, CT: Praeger.
COWARD, W. M., & SACKETT, P. R. (1990). Linearity of ability-performance relationships: A reconfirmation. Journal of Applied Psychology, 73, 297–300.
COWLES, M. (1989). Statistics in psychology: An historical perspective. Hillsdale, NJ: Erlbaum.
COX, M. V. (1993). Children's drawings of the human figure. Hove, UK: Erlbaum.
COX, R. H. (1989). Psychomotor screening for USAF pilot candidates: Selecting a valid criterion. Aviation, Space, and Environmental Medicine, 60, 1153–1156.
CRAIG, R. J. (Ed.). (1993). The Millon Clinical Multiaxial Inventory: A clinical research information synthesis. Hillsdale, NJ: Erlbaum.
CRAIK, R I. M., & SALTHOUSE, T. A. (Eds,). (1992). The handbook of aging and cognition. Hillsdale, NJ: Erlbaum.
CRAMER, P. (1996). Storytelling, narrative, and the Thematic Apperception Test. New York: Guilford Press.
CRAMER, P., & BLATT, S. J. (1990). Use of the TAT to measure change in the defense mechanisms following intensive psychotherapy. Journal of Personality Assessment, 54, 236–251.
CRAWFORD, J. E., & CRAWFORD, D. M. (1981). Crawford Small Parts Dexterity Test: Manual. San Antonio, TX: Psychological Corporation.
CRICK, G. E., & BRENNAN, R. L. (1982). GENOVA. A generalized analysis of variance system [Computer program and manual]. Dorchester: University of Massachusetts at Boston, Computer Facilities.
CRITES, J. O. (1969). The maturity of vocational attitudes in adolescence. Iowa City: University of Iowa.
CROCKER, L., & SCHMITT, A. (1987). Improving multiple-choice test performance for examinees with different levels of test anxiety. Journal of Experimental Education, 55, 201–205.
CRONBACH, L. J. (1949). Statistical methods applied to Rorschach scores: A review. Psychological Bulletin, 46, 393–429.
CRONBACH, L. J. (1951) Coefficient alpha and the internal structure of tests. Psychometrika, 16, 297–334.
CRONBACH, L. J., & DRENTH, P. J. D. (Eds.). (1972). Mental tests and cultural adaptation. The Hague: Mouton.
CRONBACH, L. J., & FURBY, L. (1970). How we should measure change — or should we? Psychological Bulletin, 74, 68–80.
CRONBACH, L. J., & GLESER, G. C. (1965). Psychological tests and personnel decisions (2nd ed.). Champaign: University of Illinois Press.
CRONBACH, L. J., GLESER, G. C., NANDA, H., & RAJARATNAM, N. (1972). The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: Wiley.
CRONBACH, L. J., & MEEHL, P. E. (1955). Construct validity in psychological tests. Psychological Bulletin, 52, 281–302.
CROWNE, D. P, & MARLOWE, D. (1964). The approval motive: Studies in evaluative dependence. New York: Wiley.
CSIKSZENTMIHALYI, M., RATHUNDE, K., & WHALEN, S. (1993). Talented teenagers: The roots of success and failure. New York: Cambridge University Press.
CUDECK, R., & O'DELL, L. L. (1994). Applications of standard error estimates in unrestricted factor analysis: Significance tests for factor loading and correlations. Psychological Bulletin, 115, 475–487.
CULBERTSON, J. L., & WILLIS, D. J. (Eds.). (1993). Testing young children: a reference guide for developmental, psychoeducational, and psychosocial assessments. Austin, TX: PRO-ED.
CULLER, R. E., & HOLAHAN, C. J. (1980). Test anxiety and academic performance: The effects of study-related behavior. Journal of Educational Psychology, 72, 16–20. Culture and psychology. (1995). Vol. 1, No. 1. Newbury Park, CA: Sage.
CUMMINGS, J. A. (1986). Projective drawings. In H. M. Knoff (Ed.), The assessment of child and adolescent personality (pp. 199–244). New York: Guilford Press.
CUNDICK, B. P. (1985). Review of the Holtzman Inkblot Technique. Ninth Mental Measurements Yearbook, Vol. 1, 661–662.
CURETON, E. E. (1950). Validity, reliability, and baloney. Educational and Psychological Measurement, 10, 94–96.
CURETON, E. E. (1957a). Recipe for a cookbook. Psychological Bulletin, 54, 494–497.
CURETON, E. E. (1957b). The upper and lower twenty-seven percent rule. Psychometrika, 22, 293–296.
CURETON, E. E. (1965). Reliability and validity: Basic assumptions and experimental designs. Educational and Psychological Measurement, 25, 327–346.
CURETON, E. E., COOK, J. A., FISCHER, R. T, LASER, S. A., ROCKWELL, N. J., & SIMMONS, J. W. (1973). Length of test and standard error of measurement. Educational and Psychological Measurement, 33, 63–68.
CUSHMAN, L. A., & SCHERER, M. J. (Eds.). (1995). Psychological assessment in medical rehabilitation. Washington, DC: American Psychological Association.
CUTTER, E., & FARBEROW, N. L. (1970). The consensus Rorschach. In B. Klopfer, M. M. Mayer, F. B. Brawer, & W. G. Klopfer (Eds.), Developments in the Rorschach technique (Vol. 3, pp. 209–261). San Diego, CA: Harcourt Brace Jovanovich.
DAHLSTROM, W. G. (1993a). The items in the MMPI-2: Alterations in wording, patterns of interrelationships, and changes in endorsements, Supplement to the MMPI-2 manual for administration and scoring. Minneapolis: University of Minnesota Press.
DAHLSTROM, W. G. (1993b). Tests: Small samples, large consequences. American Psychologist, 48, 393–399.
DAHLSTROM, W. G. (1995). Pigeons, people, and pigeon holes. Journal of Personality Assessment, 64, 2–20.
DAHLSTROM, W. G., & DAHLSTROM, L. E. (Eds.). (1980). Basic readings on the MMPI: A new selection on personality measurement. Minneapolis: University of Minnesota Press.
DAHLSTROM, W. G., & TELLEGEN, A. (1993). Socioeconomic status and the MMPI-2: The relation of MMPI-2 patterns to levels of education and occupation; Supplement to the MMPI-2 manual for administration and scoring. Minneapolis: University of Minnesota Press.

DAHLSTROM, W. G., WELSH, G. S., & DAHLSTROM, L. E. (1972). An MMPI handbook: Vol. 1. Clinical interpretation. Minneapolis: University of Minnesota Press.
DAHLSTROM, W. G., WELSH, G. S., & DAHLSTROM, L. E. (1975). An MMPI handbook: Vol. 2. Research applications. Minneapolis: University of Minnesota Press.
DANA, R. H. (1984). Intelligence testing of American Indian children: Sidesteps in quest of ethnical practice. White Cloud Journal, 3 (3), 35–43.
DANA, R. H. (1993). Multicultural assessment perspectives for professional psychology. Boston: Allyn & Bacon. DANA, R. H. (1996a). Culturally competent assessment practice in the United States. Journal of Personality Assessment, 66, 472–487.
DANA, R. H. (1996b). The Thematic Apperception Test (TAT). In C. S. Newmark (Ed.), Major psychological assessment instruments (2nd ed., pp. 166–205). Boston: Allyn & Bacon.
DANIELS, D., & PLOMIN, R. (1985). Differential experience of siblings in the same family. Developmental Psychology, 21, 747–760.
DANIELS, M. H. (1989). Review of the Self-Directed Search: A guide to educational and vocational planning — 1985 Revision. Tenth Mental Measurements Yearbook, 735–738.
DARLINGTON, R. B. (1971). Another look at «culture fairness.» Journal of Educational Measurement, 8, 71–82.
DARLINGTON, R. B. (1976). A defense of «rational» personnel selection, and two new methods. Journal of Educational Measurement, 13, 43–52.
DARLINGTON, R. B., & STAUFFER, G. F. (1966). A method for choosing a cutting point on a test. Journal of Applied Psychology, 50, 229–231.
DAS, J. P. (1984). Simultaneous and successive processes and K-ABC. Journal of Special Education, 18, 229–238.
DAS, J. P., KIRBY, J. R., & JARMAN, R. F. (1975). Simultaneous and successive syntheses: An alternative model for cognitive abilities. Psychological Bulletin, 82, 87–103.
DAS, J. P., KIRBY, J. R., & JARMAN, R. F. (1979). Simultaneous and successive cognitive processes. New York: Academic Press.
DAS, J. P., & MOLLOY, G. N. (1975). Varieties of simultaneous and successive processing in children. Journal of Educational Psychology, 67, 213–220.
DAS, J. R., NAGLIERI, J. A., & KIRBY, J. R. (1994). Assessment of cognitive processes: The PASS theory of intelligence. Boston: Allyn & Bacon.
DAS, R. S. (1963). Analysis of the components of reasoning in nonverbal tests and the structure of reasoning in a bilingual population. Archiv fur die Gesamte Psychologies, 115 (3), 217–229.
DASEN, P. K. (Ed.). (1977). Piagetian psychology: Cross-cultural contributions. New York: Halsted Press.
DAVIDOW, S., & BRUHN, A. R. (1990). Earliest memories and the dynamics of delinquency: A replication study. Journal of Personality Assessment 54, 601–616.
DAVIS, C. J. (1980). Perkins-Binet Tests of Intelligence for the Blind. Watertown, MA: Perkins School for the Blind.
DAVIS, D. L., GROVE, S, J., & KNOWLES, P. A. (1990). An experimental application of personality type as an analogue for decision-making style. Psychological Reports, 66, 167–175.
DAVIS, F. B. (1959). Interpretation of differences among averages and individual test scores. Journal of Educational Psychology, 50, 162–170.
DAVIS, G. L., HOFFMAN, R. G., & NELSON, K. S. (1990). Differences between Native Americans and Whites on the California Psychological Inventory. Psychological Assessment, 2, 238–242.
DAVIS, W. E. (1969a). Effect of prior failure on subjects' WAIS Arithmetic subtest scores. Journal of Clinical Psychology, 25, 72–73.
DAVIS, W. E. (1969b). Examiner differences, prior failure, and subjects' arithmetic scores. Journal of Clinical Psychology, 25, 178–180.
DAVISON, M. L., GASSER, M., & DING, S. (1996). Identifying major profile patterns in a population: An exploratory study of WAIS and GATB patterns. Psychological Assessment, 8, 26–31.
DAWES, R. M., FAUST, D., & MEEHL, P. E. (1993). Statistical prediction versus clinical prediction: Improving what works. In G. Keren & C. Lewis (Eds.), A handbook for data analysis in the behavioral sciences: Methodological issues (pp. 351–367). Hillsdale, NJ: Erlbaum.
DAWIS, R. V. (1991). Vocational interests, values, and preferences. In M. D. Dunnette & L. M. Hough (Eds.), Handbook of industrial and organizational psychology (2nd ed., Vol. 2, pp. 833–871). Palo Alto, CA: Consulting Psychologists Press.
DAWIS, R. V. (1992). The structure(s) of occupations: Beyond RIASEC. Journal of Vocational Behavior, 40, 171–178.
DEAN, R. S. (1977). Reliability of the WISC-R with Mexican-American children. Journal of School Psychology, 15, 267–268.
DEAN, R. S. (1979). Predictive validity of the WISC-R with Mexican-American children. Journal of School Psychology, 17, 55–58.
DEAN, R. S. (1980). Factor structure of the WISC-R with Anglos and Mexican-Americans. Journal of School Psychology, 18, 234–239.
DEAN, R. S. (1985). Review of Halstead-Reitan Neuropsychological Test Battery. Ninth Mental Measurements Yearbook, Vol. 1, 644–646.
DE GROOT, A. M. B., & BARRY, C. (Eds.). (1993). The multilingual community: Bilingualism. European Journal of Cognitive Psychology, 4(4). Hove, England: Erlbaum.
DEKKER, R. (1993). Visually impaired children and haptic intelligence test scores: Intelligence Test for Visually Impaired Children (ITVIC). Developmental Medicine and Child Neurology, 35, 478–489.
DEKKER, R., DRENTH, P. J. D., & ZAAL, J. N. (1991). Results of the Intelligence Test for Visually Impaired Children (ITVIC). Journal of Visual Impairment and Blindness, 85, 261–267.
DEKKER, R., DRENTH, R J. D., ZAAL, J. N., & KOOLE, P. D. (1990). An intelligence test series for blind and low vision children. Journal of Visual Impairment and Blindness, 84, 71–76.
DEKKER, R., & KOOLE, P. D. (1992). Visually impaired children's visual characteristics and intelligence. Developmental Medicine and Child Neurology, 34, 123–133.

DELANEY, E. & HOPKINS, T. (1987). Stanford-Binet Intelligence Scale — Examiners hand-book: An expanded guide for fourth edition users. Chicago: Riverside.
DeMERS, S. T., FIORELLO, C., & LANGER, K. L. (1992). Legal and ethical issues in preschool assessment. In E. Vazquez Nutall, I. Romero, & J. Kalesnik (Eds.), Assessing and screening preschoolers: Psychological and educational dimensions (pp. 43–54). Boston: Allyn & Bacon.
DEMETRIOU, A. (1988). The Neo-Piagetian theories of cognitive development: Toward an integration. Amsterdam: North-Holland.
DEMMING, J. A., & PRESSEY, S. L. (1957). Tests «indigenous to the adult and older years. Journal of Counseling Psychology, 4, 144–148.
DEMO, D. H. (1985). The measurement of self-esteem: Refining our methods. Journal of Personality and Social Psychology, 48, 1490–1502.
DENNIS, W. (1966). Goodenough scores, art experience, and modernization. Journal of Social Psychology, 68, 211–228.
DENNY, J. P. (1966). Effects of anxiety and intelligence on concept formation. Journal of Experimental Psychology, 72, 596–602.
DENO, S. L. (1992). The nature and development of curriculum-based measurement. Presenting School Failure, 36, 5–10.
DEPAULO, B. M. (1994). Spotting lies: Can humans learn to do better? Current Directions in Psychological Science, 3, 83–86.
DEROGATIS, L. R. (1994). SCL–90-R: Symptom Checklist–90-R: Administration, scoring, and procedures manual (3rd ed.). Minneapolis, MN: National Computer Systems.
DEROGATIS, L. R., & LAZARUS, L. (1994). SCL–90-R, Brief Symptom Inventory, and matching clinical rating scales. In M. E. Maruish (Ed.), The use of psychological testing for treatment planning and outcome assessment (pp. 217–248). Hillsdale, NJ: Erlbaum.
DESMARAIS, L. B., MASI, D. L., OLSON, M. J., BARBERA, K. M., & DYER, P. J. (1994, April). Scoring a multimedia situational Judgment test: JBMs experience. Paper presented at the annual conference of the Society for Industrial and Organizational Psychology, Nashville, TN.
DETTERMAN, D. K. (Ed.). (1985–1993). Current topics in human intelligence (Vols. 1–3). Norwood, NJ: Ablex.
DETTERMAN, D. K., & STERNBERG, R. J. (Eds.). (1982). How and how much can intelligence be increased. Norwood, NJ: Ablex.
DeVITO, A. J. (1985). Review of Myers-Briggs Type Indicator. Ninth Mental Measurements Yearbook, Vol. 2, 1030–1032.
DeWITT, L. J., & WEISS, D. J. (1974). Computer software system for adaptive ability measurement (Res. Rep. 74–1). Minneapolis: Department of Psychology, University of Minnesota, Psychometric Methods Program.
DeWOLFF, C. J. (1993). The prediction paradigm. In H. Schuler, J. L. Farr, & M. Smith (Eds.), Personnel selection and assessment: Individual and organizational perspectives (pp. 253–265). Hillsdale, NJ: Erlbaum.
DIANE, C. C., BROGAN, E S., & McCAULEY, D. E., JR. (1991). A validation study of artificial language tests for border patrol guards. Washington, DC: Office of Personnel Research and Development.
DIAZ-GUERRERO, R. (1990). The need for ethnopsychology of cognition and personality. In I. Ayman & Y. Tanaka (Organizers) Symposium: Appropriate Psychology for developing countries Kyoto, Japan: International Association of Applied Psychology for Developing Countries.
DIAZ-GUERRERO, R., & DIAZ-LOVING, R. (1990). Interpretation in cross-cultural personality assessment. In C. R. Reynolds & R. W. Kamphaus (Eds.), Handbook of psychological and educational assessment of children: Personality, behavior and context (pp. 491–523). New York: Guilford Press.
DIAZ-GUERRERO, R., & SZALAY, L. B. (1991). Understanding Mexicans and Americans, New York: Plenum Press.
DICKINSON, T. L., & ZELLINGER, P. M. (1980). A comparison of behaviorally anchored rating and mixed standard scale formats. Journal of Applied Psychology, 65, 147–154.
DIENER, E., & CRANDALL, R. (1978). Ethics in social and behavioral research, Chicago: University of Chicago Press.
Differential Aptitude Tests, Fifth Edition: Counselor's Manual. (1991). San Antonio, TX: Psychological Corporation.
Differential Aptitude Tests, Fifth Edition: Technical Manual. (1992). San Antonio, TX: Psychological Corporation.
DIGMAN, J. M. (1990). Personality structure: Emergence of the Five-Factor Model. Annual Review of Psychology, 41, 417–440.
DOBBIN, J. (1984). How to take a test: Doing your best. Princeton, NJ: Educational Testing Service.
DOLL, E. A. (1965). Vineland Social Maturity Scale: Manual of directions (rev. ed.). Circle Pines, MN: American Guidance Service. (1st ed., 1935)
DONDERS, J. (1996). Cluster subtypes in the WISC-III standardization sample: Analysis of factor index scores. Psychological Assessment, 8, 32 2–318.
DONLON, T. F. (Ed.). (1984). The College Board technical handbook for the Scholastic Aptitude Test and achievement tests. New York: College Board Publications.
DOYLE, K. O., JR. (1974). Theory and practice of ability testing in ancient Greece. Journal of the History of the Behavioral Sciences, 10, 202–212.
DRASGOW, F., & HULIN, C. L. (1990). Item response theory. In M. D. Dunnette & L. M. Hough (Eds.), Handbook of industrial and organizational psychology (2nd ed., Vol. 1, pp. 577–636). Palo Alto, CA: Consulting Psychologists Press.
DRASGOW, F., OLSON-BUCHANAN, J. B., & MOBERG, P. J. (1996). Development of interactive video assessments. Manuscript submitted for publication.
DREGER, R. M. (1968). General temperament and personality factors related to intellectual performances. Journal of Genetic Psychology, 113, 275–293.
DROEGE, R. C. (1966). Effects of practice on aptitude scores. Journal of Applied Psychology, 50, 306–310.
DRUMMOND, R. J. (1995). Review of the Alcohol Use Inventory. Twelfth Mental Measurements Yearbook, 65–66.
DRUMMOND, R. J. (1996). Appraisal procedures for counselors and helping professionals (3rd ed.). Englewood Cliffs, NJ: Merrill.
DUBOIS, P. H. (1939). A test standardized on Pueblo Indian children. Psychological Bulletin, 36,523.
DUBOIS, P. H. (1970). A history of psychological testing. Boston: Allyn & Bacon.
DUCKWORTH, J. C. (1991). The Minnesota Multiphasic Personality Inventory–2: A review. Journal of Counseling and Development 69, 564–567.

DUDEK, F. J. (1979). The continuing misinterpretation of the standard error of measurement. Psychological Bulletin, 86, 335–337.
DUNCAN, O. D. (1961). A socioeconomic index for all occupations. In A. J. Reiss, Jr. (Ed.), Occupations and social status (pp. 109–138). New York: Free Press of Glencoe.
DUNN, J. A. (1967). Inter- and intra-rater reliability of the new Harris-Goodenough Draw-a-Man Test. Perceptual and Motor Skills, 24, 269–270.
DUNN, J., & PLOMIN, R. (1990). Separate lives: Why siblings are so different. New York: Basic Books.
DUNN, L(LOYD), M., & DUNN, L(EOTA), M. (1981). Peabody Picture Vocabulary Test-Revised: Manual for Forms L and M. Circle Pines, MN: American Guidance Service.
DUNNETTE, M. D. (1957). Use of the sugar pill by industrial psychologists. American Psychologist, 12, 223–225.
DUNNETTE, M. D., & BORMAN, W. C. (1979). Personnel selection and classification systems. Annual Review of Psychology, 30, 477–525.
DUNNETTE, M. D., & HOUGH, L. M. (Eds.). (1990–1992). Handbook of industrial and organizational psychology (2nd ed., Vols. 1–3). Palo Alto, CA: Consulting Psychologists Press.
DUNST, C. J. (1980). A clinical and educational manual for use with the Užgiris and Hunt Scales of Infant Psychological Development. Austin, TX: PRO-ED.
DUNST, C. J., & GALLAGHER, J. L. (1983). Piagetian approaches to infant assessment. Topics in Early Childhood Special Education, 3, 44–62.
DURAN, R. P. (1983). Hispanics' education and background: Predictors of college achievement. New York: College Entrance Examination Board.
DURAN, R. P. (1989). Testing of linguistic minorities. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 573–587). New York: American Council on Education/ Macmillan.
DUSH, D. M. (1985). Review of the Holtzman Inkblot Technique. Ninth Mental Measurements Yearbook, Vol. 1, 602–603.
DWYER, C. A. (1993). Innovation and reform: Examples from teacher assessment. In R. E. Bennett & W. C. Ward (Eds.), Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment (pp. 265–289). Hillsdale, NJ: Erlbaum.
DYER, H. S. (1973). Recycling the problems of testing. Proceedings of the 1972 Invitational Conference on Testing Problems, Educational Testing Service, 85–95.
EAGLY, A. H., & CHAIKEN, S. (1993). The psychology of attitudes. Fort Worth, TX: Harcourt Brace Jovanovich.
EBBINGHAUS, H. (1897). Uber eine neue Methode zur Pruning geistiger Fahigkeiten und ihre Anwendung bei Schulkindern. Zeitschrift fur Angewandte Psychologie, 13, 401–459.
EBEL, R. L. (1962). Content standard test scores. Educational and Psychological Measurement, 22, 15–25.
EBEL, R. L. (1972). Some limitations of criterion-referenced measurement. In G. H. Bracht, K. D. Hopkins, & J. C. Stanley (Eds.), Perspective in educational and psychological measurement (pp. 144–149). Englewood Cliffs, NJ: Prentice Hall.
EBEL, R. L. (1979). Essentials of educational measurement (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.
EBEL, R. L., & DAMRIN, D. E. (1960). Tests and examinations. Encyclopedia of educational research (3rd ed., pp. 1502–1517). New York: Macmillan.
EDER, R. W., KACMAR, K. M., & FERRIS, G. R. (1989). Employment interview research: History and synthesis. In R. W. Eder & G. R. Ferris (Eds.), The employment interview: Theory, research, and practice (pp. 17–31). Newbury Park, CA: Sage.
EDUCATIONAL TESTING SERVICE. (1990). Annual report. Princeton, NJ: Author.
EDUCATIONAL TESTING SERVICE. (1992). ETS conference examines the technology of computer-based testing for people with disabilities. ETS Developments, 38(1), 6–7.
EDWARDS, A. L. (1957). The social desirability variable in personality assessment and research. New York: Dryden.
EDWARDS, A. L. (1959). Edwards Personal Preference Schedule: Manual. New York: Psychological Corporation.
EDWARDS, A. L. (1990). Construct validity and social desirability. American Psychologist, 45, 287–289.
EICHORN, D. H., CLAUSEN, J. A., HAAN, N., HONZIK, M. P., & MUSSEN, P. H. (Eds.). (1981). Present and past in middle life. New York: Academic Press.
The Eighth Mental Measurements Yearbook. (1978). Highland Park, NJ: Gryphon Press.
EISDORFER, C. (1963). The WAIS performance of the aged: A retest evaluation. Journal of Gerontology, 18, 169–172.
EKSTROM, R. B., FRENCH, J. W., & HARMAN, H. H. (1979). Cognitive factors: Their identification and replication. Multivariate Behavioral Research Monographs, No. 79–2.
EKSTROM, R. B., FRENCH, J. W., HARMAN, H. H., & DERMEN, D. (1976). Manual for kit of factor-referenced cognitive tests (3rd ed.). Princeton, NJ: Educational Testing Service.
The Eleventh Mental Measurements Yearbook. (1992). Lincoln, NE: Buros Institute of Mental Measurements.
ELKSNIN, L. K., & ELKSNIN, N. (1993). A review of picture interest inventories: Implications for vocational assessment of students with disabilities. Journal of Psychoeducational Assessment, 11, 323–336.
ELLIOTT, C. D. (1990a). Differential Ability Scales: Administration and scoring manual. San Antonio, TX: Psychological Corporation.
ELLIOTT, C. D. (1990b). Differential Ability Scales: Introductory and technical handbook. San Antonio, TX: Psychological Corporation.
ELLIOTT, C. D., MURRAY, D. J., & PEARSON, L. S. (1979), British Ability Scales. Windsor, England: National Foundation for Educational Research.
EMBRETSON, S. E. (1983). Construct validity: Construct representation versus nomothetic span. Psychological Bulletin, 93, 179–197.
EMBRETSON, S. E. (Ed.). (1985a). Test design: Developments in psychology and psychometrics. Orlando, FL: Academic Press.
EMBRETSON, S. E. (1985b). Multicomponent latent trait models for test design. In S. E. Embretson (Ed.), Test design: Developments in psychology and psychometrics (pp. 195–218). Orlando, FL: Academic Press.
EMBRETSON, S. E. (1986). Intelligence and its measurement: Extending contemporary theory to existing tests. In R. J. Sternberg (Ed.), Advances in the psychology of human intelligence (Vol. 3, pp. 335–368). Hillsdale, NJ: Erlbaum.
EMBRETSON, S. E. (1987). Toward development of a psychometric approach. In C. S. Lidz (Ed.), Dynamic assessment: An interactive approach to evaluating learning potential (pp. 135–164). New York: Guilford Press.

EMBRETSON, S. (1990). *Diagnostic testing by measuring learning processes: Psychometric considerations for dynamic testing.* In N. Frederiksen, R. Glaser, A. Lesgold, & M. G. Shafto (Eds.), Diagnostic monitoring of skill and knowledge acquisition (pp. 407–432). Hillsdale, NJ: Erlbaum.
EMBRETSON, S. E. (1991). A multidimensional latent trait model for measuring learning and change. Psychometrika, 56, 495–515.
EMBRETSON, S. E. (1992). Computerized adaptive testing: Its potential substantive contributions to psychological research and assessment. Current Directions in Psychological Science, 1,129–131.
EMBRETSON, S. (1993). Psychometric models for learning and cognitive processes. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), Test theory for a new generation of tests (pp. 125–150). Hillsdale, NJ: Erlbaum.
EMBRETSON, S. (1994). Applications of cognitive design systems to test development. In C. R. Reynolds (Ed.), Cognitive assessment: A multidisciplinary perspective (pp. 107–135). New York: Plenum Press.
EMBRETSON, S. E. (1995a). Developments toward a cognitive design system for psychological tests. In D. Lubinsky & R. V. Dawis (Eds.), Assessing individual differences in human behavior: New methods, concepts, and findings (pp. 17–46). Palo Alto, CA: Consulting Psychologists Press.
EMBRETSON, S. E. (1995b). A measurement model for linking individual learning to processes and knowledge: Application to mathematical reasoning. Journal of Educational Measurement, 32, 277–294.
Encyclopedia of human intelligence. (1994). New York: Macmillan.
ENDLER, N. S., & MAGNUSSON, D. (1976). Toward an interactional psychology of personality. Psychological Bulletin, 83, 956–974.
ENGELHARD, G. (1992). Review of the California Psychological Inventory, Revised Edition. Eleventh Mental Measurements Yearbook, 139–141.
ENGELHART, M. D. (1965). A comparison of several item discrimination indices. Journal of Educational Measurement, 2, 69–76.
ENTWISLE, D. R. (1972). To dispel fantasies about fantasy-based measures of achievement motivation. Psychological Bulletin, 77, 377–391.
EPSTEIN, S. (1966). Some theoretical considerations on the .nature of ambiguity and the use of stimulus dimensions in projective techniques. Journal of Counseling Psychology, 30,183–192.
EPSTEIN, S. (1979). The stability of behavior: I. On predicting most of the people much of the time. Journal of Personality and *Social Psychology, 37, 1097–1121.*
EPSTEIN, S. (1980). The stability of behavior: II. Implications for psychological research. American Psychologist, 35, 790–806.
EPSTEIN, S., & O'BRIEN, E. J. (1985). The person-situation debate in historical and current perspective. Psychological Bulletin, 98, 513–537.
EPTING, E., & LANDFIELD, A. W. (Eds.). (1985). Anticipating personal construct psychology. Lincoln: University of Nebraska Press.
EQUAL EMPLOYMENT OPPORTUNITY COMMISSION. (EEOC). (1994, May). Enforcement guidance: Preemployment disability-related inquiries and medical examinations under the Americans with Disabilities Act of 1990 (EEOC Notice, 915.002). Washington, DC: Author.
EQUAL EMPLOYMENT OPPORTUNITY COMMISSION (EEOC). (1995, October). ADA enforcement guidance: Preemployment disability-related questions and medical examinations. Washington, DC: Author.
ERDBERG, P., & EXNER, J. E., JR. (1984). Rorschach assessment. In G. Goldstein & M. Hersen (Eds.), Handbook of psychological assessment (pp. 332–347). New York: Pergamon.
ERICSSON, K. A. (1987). Theoretical implications from protocol analysis on testing and measurement. In R. R. Ronning, J. A. Glover, J. C. Conoley, & J. C. Witt (Eds.), The influence of cognitive psychology on testing (pp. 191–226). Hillsdale, NJ: Erlbaum.
ERICSSON, K. A., & SIMON, H. A. (1993). Protocol analysis: Verbal reports as data (rev. ed.). Cambridge, MA: MIT Press.
ERICSSON, K. A., & SMITH, J. (Eds.). (1991). Toward a general theory of expertise: Prospects and limits. New York: Cambridge University Press.
ESQUIROL, J. E. D. (1838). Des maladies mentales considerees sous les rapports medical, hygienique, et medico-legal (2 vols».). Paris: Bailliere.
ESTES, W. K. (1974). Learning theory and intelligence. American Psychologist, 29, 740–749.
ETS kit of factor-referenced cognitive tests. (1976). Princeton, NJ: Educational Testing Service.
ETS Standards for quality and fairness. (1987). Princeton, NJ: Educational Testing Service. (Original edition published 1981)
EVANS, F. R., & PIKE, L. W. (1973). The effects of instruction for three mathematics item formats. Journal of Educational Measurement, 20,25 7–2 72.
EXNER, J. E., JR. (1966). Variations in WISC performances as influenced by differences in pretest rapport. Journal of General Psychology, 74, 299–306.
EXNER, J. E., JR. (1969). The Rorschach systems. New York: Grune & Stratton.
EXNER, J. E., JR. (1974). The Rorschach: A comprehensive system. New York: Wiley.
EXNER, J. E., JR. (1989). Searching for projection in the Rorschach. Journal of Personality Assessment, 53, 520–536.
EXNER, J. E., JR. (1991). The Rorschach: A comprehensive system: Vol. 2. Interpretation (2nd ed.). New York: Wiley.
EXNER, J. E., JR.(1992). R in Rorschach research: A ghost revisited. Journal of Personality Assessment, 58, 245–251.
EXNER, J. E., JR. (1993). The Rorschach: A comprehensive system: Vol. 1. Basic foundations (3rd ed.). New York: Wiley.
EXNER, J. E., JR. (Ed.). (1995). Issues and methods in Rorschach research. Mahwah, NJ: Erlbaum.
EXNER, J. E., JR. (1996). A comment on «The Comprehensive System for the Rorschach: A critical examination.» Psychological Science, 7, 11–13.
EXNER, J. E., JR., & WEINER, I. B. (1995). The Rorschach: A comprehensive system: Vol. 3. Assessment of children and adolescents (2nd ed.). New York: Wiley.
EYDE, L. D. (1987). Computerized psychological testing: An introduction. Applied Psychology: An International Review, 36(3/4), 223–235.
EYDE, L. D., & KOWAL, D. M. (1987). Computerized test interpretation services: Ethical and professional concerns regarding U.S. producers and users. Applied Psychology: An International Review, 36(3/4), 401–417.

EYDE, L. D., MORELAND, K. L., ROBERTSON, G. J., PRIMOFF, E. S., & MOST, R. B. (1988). Test User Qualifications: A data-based approach to promoting good test use. Issues in scientific psychology. Washington, DC: American Psychological Association, Science Directorate.
EYDE, L. D., NESTER, M. A., HEATON, S. M., & NELSON, A. V. (1994). Guide for administering written employment examinations to persons with disabilities. Washington, DC: U. S. Office of Personnel Management.
EYDE, L. D., & QUAINTANCE, M. K. (1988). Ethical issues and cases in the practice of personnel psychology. Professional Psychology: Research and Practice, 19(2), 148–154.
EYDE, L. D., ROBERTSON, G. J., KRUG, S. E., MORELAND, K. L., ROBERTSON, A. G., SHEWAN, C. M., HARRISON, P. L., PORCH, B. E., HAMMER, A. L., & PRIMOFF, E. S. (1993). Responsible test use: Case studies for assessing human behavior. Washington, DC: American Psychological Association.
FAGAN, J. P. (1992). Intelligence: A theoretical viewpoint. Current Directions in Psychological Science, 1, 82–86.
FAGAN, J. P., & DETTERMAN, D. K. (1992). The Fagan Test of Infant Intelligence: A technical summary. Journal of Applied Developmental Psychology, 13, 173–193.
FAGGEN, J. (1987). Golden Rule revisited: Introduction. Educational Measurement: Issues and Practice, 6, 5–8
FANTUZZO, J. W., BLAKEY, W. A., & GORSUCH, R. L. (1989). WAIS-R: Administration and scoring training manual. San Antonio, TX: Psychological Corporation.
FARR, J. M. (Ed.). (1992). The complete guide for occupational exploration. Indianapolis, IN: JIST.
FEAGANS, L. V., SHORT, E. J., & MELTZER, L. J. (Eds.). (1991). Subtypes of learning disabilities: Theoretical perspectives and research. Hillsdale, NJ: Erlbaum.
FEAR, R. A., & CHIRON, R. J. (1990). The evaluation interview (4th ed.). New York: McGraw-Hill.
FEDERAL REGISTER. (1977). Handicapped Children Rule, 42(250). Washington, DC: U.S. Government Printing Office.
FEDORAK, S., & COLES, E. M. (1979). Ipsative vs. normative interpretation of test scores: A comment on Alien and Forman's (1976) norms on Edwards Personal Preference Schedule for female Australian therapy students. Perceptual and Motor Skills, 48, 919–922.
FEINGOLD, A. (1995). The additive effects of differences in central tendency and variability are important in comparisons between groups. American Psychologist, 50, 5–13.
FELDHUSEN, J. E, & KLAUSMEIER, H. J. (1962). Anxiety, intelligence, and achievement in children of low, average, and high intelligence. Child Development, 33, 403–409.
FELDMAN, D. H., & BRATTON, J. C. (1972). Relativity and giftedness: Implications for equality of educational opportunity. Exceptional Children, 38, 491–492.
FELDMAN, J. M. (1986). Instrumentation and training for performance appraisal: A perceptual-cognitive viewpoint. In K. M. Rowland & G. Ferris (Eds.), Research in personnel and human resources management (Vol. 4). Greenwich, CT: JAI Press.
FELDT, L. S., & BRENNAN, R. L. (1989). Reliability. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 105–146). New York: American Council on Education/ Macmillan.
FERGUSON, G. A. (1954). On learning and human ability. Canadian Journal of Psychology 8, 95–112.
FERGUSON, G. A. (1956). On transfer and the abilities of man. Canadian Journal of Psychology, 10,121–131.
FERGUSON, R. L., & NOVICK, M. R. (1973). Implementation of a Bayesian system for decision analysis in a program of individually prescribed instruction (ACT Res. Rep. No. 60). Iowa City: American College Testing Program.
FEUER, M. J., & KOBER, N. (Eds.). (1995). Anticipating Goals 2000: Standards, assessment, and public policy. Washington, DC: National Academy Press.
FEUERSTEIN, R. (1979). The dynamic assessment of retarded performers: The Learning Potential Assessment Device, theory, instruments, and techniques. Baltimore: University Park Press.
FEUERSTEIN, R. (1980). Instrumental enrichment: An intervention program for cognitive modifiability. Baltimore: University Park Press.
FEUERSTEIN, R. (1991). Cultural difference and cultural deprivation: Differential patterns of adaptability. In N. Bleichrodt & P. J. D. Drenth (Eds.), Contemporary issues in cross-cultural psychology (pp. 21–33). Amsterdam: Swets & Zeitlinger.
FEUERSTEIN, R., & FEUERSTEIN, S. (1991). Mediated learning experience: A theoretical review. In R. Feuerstein, P. S. Klein, & A. J. Tannenbaum (Eds.), Mediated learning experience (MLE): Theoretical, psychosocial, and learning implications (pp. 3–51). London: Freund.
FEUERSTEIN, R., RAND, Y., JENSEN, M. R., KANIEL, S., & TZURIEL, D. (1987). Prerequisites for assessment of learning potential: The LPAD model. In C. S. Lidz (Ed.), Dynamic assessment: An interactive approach to evaluating learning potential (pp. 35–51). New York: Guilford Press.
FEWELL, R. R. (1991). Assessment of visual functioning. In B. A. Bracken (Ed.), The psychoeducational assessment of preschool children (2nd ed., pp. 317–340). Boston: Allyn & Bacon.
FIGUEROA, R. A. (1990). Assessment of linguistic minority group children. In C. R. Reynolds & R. W. Kamphaus (Eds.), Handbook of psychological and educational assessment of children: Intelligence and achievement (pp. 671–696). New York: Guilford Press.
FIGURELLI, J. C., & KELLER, H. R. (1972). The effects of training and socioeconomic class upon the acquisition of conservation concepts. Child Development, 43, 293–298.
Finding information about psychological tests. (1995). Washington, DC: American Psychological Association, Science Directorate.
FINK, A. (Ed.). (1995). The survey kit (Vols. 1–9). Thousand Oaks, CA: Sage.
FINKLE, R. B. (1983). Managerial assessment centers. In M. D. Dunnette (Ed.), Handbook of industrial and organizational psychology (pp. 861–888). New York: Wiley.
FISCHER, C. T. (1985). Individualizing psychological assessment. Monterey, CA: Brooks/Cole.
FISCHER, J., & CORCORAN, K. (1994). Measures for clinical practice: A sourcebook (2nd ed., vols. 1–2). New York: Free Press.
FISKE, D. W. (1949). Consistency of the factorial structures of personality ratings from different sources. Journal of Abnormal and Social Psychology, 44, 329–344.

FISKE, D. W. (1973). Can a personality construct be validated empirically? Psychological Bulletin, 80, 89–92.
FISKE, M., & CHIRIBOGA, D. A. (1990). Change and continuity in adult life. San Francisco: Jossey-Bass.
FITZGERALD, B. J., PASEWARK, R. A., & FLEISHER, S. (1974). Responses of an aged population on the Gerontological and Thematic Apperception Tests. Journal of Personality Assessment, 38, 234–235.
FITZMAURICE, C., & WITT, J. C. (1989). Review of the Boehm Test of Basic Concepts-Revised. Tenth Mental Measurements Yearbook, 101–102.
FLANAGAN, D. P., & ALFONSO, V. C. (1995). A critical review of the technical characteristics of new and recently revised intelligence tests for preschool children. Journal of Psychoeducational Assessment, 13, 66–90.
FLANAGAN, D. P., GENSHAFT, J. L., & HARRISON, P. L. (Eds.). (1997). Contemporary intellectual assessment: Theories, tests, and issues. New York: Guilford Press.
FLANAGAN, J. C. (1947). Scientific development of the use of human resources: Progress in the Army Air Forces. Science, 105, 57–60.
FLANAGAN, J. C. (1949). Critical requirements: A new approach to employee evaluation. Personnel Psychology, 2, 419–425.
FLANAGAN, J. C. (1954). The critical incident technique. Psychological Bulletin, 51, 327–358.
FLANAGAN, J. Q (1962). Symposium: Standard scores for aptitude and achievement tests: Discussion. Educational and Psychological Measurement, 22, 35–39.
FLAVELL, J. H. (1963). The developmental psychology of Jean Piaget. New York: Van Nostrand-Reinhold.
FLAVELL, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. American Psychologist, 34, 906–911.
FLEISHMAN, E. A. (1972). On the relation between abilities, learning, and human performance. American Psychologist, 27, 1018–1032.
FLEISHMAN, E. A. (1975). Toward a taxonomy of human performance. American Psychologist, 30, 1121–1149.
FLEISHMAN, E. A., & MUMFORD, M. D. (1989). Abilities as causes of individual differences in skill acquisition. Human Performance, 2 (3), 201–223.
FLEISHMAN, E. A., & MUMFORD, M. D. (1991). Evaluating classifications of job behavior: A construct validation of the ability requirement scales. Personnel Psychology, 44, 523–575.
FLEISHMAN, E. A., & QUAINTANCE, M. K. (1984). Taxonomies of human performance: The description of human tasks. Orlando, PL: Academic Press.
FLEISHMAN, E. A., & REILLY, M. E. (1992a). Administrators guide F-JAS: Fleishman Job Analysis Survey. Bethesda, MD: Management Research Institute.
FLEISHMAN, E. A., & REILLY, M. E. (1992b). Handbook of human abilities: Definitions, measurements, and job task requirements. Bethesda, MD: Management Research Institute.
FLEMING, J. S., & COURTNEY, B. E. (1984). The dimensionality of self-esteem. II. Hierarchical facet model for revised measurement scales. Journal of Personality and Social Psychology, 46, 404–421.
FLEMING, J. S., & WHALEN, D. J. (1990). The Personal and Academic Self-Concept Inventory: Factor structure and gender differences in high school and college samples. Educational and Psychological Measurement, 50, 957–967.
FLYNN, J. R. (1984). The mean IQ of Americans: Massive gains 1932 to 1978. Psychological Bulletin, 95, 29–51.
FLYNN, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. Psychological Bulletin, 101, 171–191.
FOOTE, J., & KAHN, M. W. (1979). Discriminative effectiveness of the Senior Apperception Test with impaired and nonimpaired elderly persons. Journal of Personality Assessment, 43, 360–364.
FORSTER, A. A., & MATARAZZO, J. D. (1990). Assessing the intelligence of adolescents with the Wechsler Adult Intelligence Scale-Revised (WAIS-R). In C. R. Reynolds & R. W. Kamphaus (Eds.), Handbook of psychological and educational assessment of children (pp. 166–182). New York: Guilford Press.
FORTIER, M. S., VALLERAND, R. J., & GUAY, F. (1995). Academic motivation and school performance: Toward a structural model. Contemporary Educational Psychology, 20, 257–274.
FOUAD, N. A., & DANCER, L. S. (1992). Cross-cultural structure of interests: Mexico and the United States. Journal of Vocational Behavior, 40, 129–143.
FOWLER, R. D., & BUTCHER, J. N. (1986). Critique of Matarazzo's views on computerized testing: All sigma and no meaning. American Psychologist, 41, 94–96.
FOX, R. A., & MEYER, D. J. (1990). Assessment of adaptive behavior. In A. F. Rotatori, R. A. Fox, D. Sexton, & J. Miller (Eds.), Comprehensive assessment in special education: Approaches, procedures, and concerns (pp. 309–338). Springfield, IL: Charles C Thomas.
FRANSELLA, R., & THOMAS, L. (1988). Experimenting with personal construct psychology. London: Routledge, Chapman & Hall.
FRANZ, S. I. (1919). Handbook of mental examination methods (2nd ed.). New York: Macmillan.
FREDERIKSEN, C. H. (1969). Abilities, transfer, and information retrieval in verbal learning. Multivariate Behavioral Research Monographs, No. 69–2.
FREDERIKSEN, N. (1962). Factors in in-basket performance. Psychological Monographs, 76(22, Whole No. 541).
FREDERIKSEN, N. (1965). Response set scores as predictors of performance. Personnel Psychology, 18, 225–244.
FREDERIKSEN, N. (1966). In-basket tests and factors in administrative performance. In A. Anastasi (Ed.), Testing problems in perspective (pp. 208–221). Washington, DC: American Council on Education.
FREDERIKSEN, N., & GILBERT, A. C. F. (1960). Replication of a study of differential predictability. Educational and Psychological Measurement, 20, 759–767.
FREDERIKSEN, N., & MELVILLE, S. D. (1954). Differential predictability in the use of test scores. Educational and Psychological Measurement, 14, 647–656.
FREEBERG, N. E. (1969). Relevance of rater-ratee acquaintance in the validity and reliability of ratings. Journal of Applied Psychology, 53, 518–524.
FREEDENFELD, R. N., ORNDUFF, S. R., & KELSEY, R. M. (1995). Object relations and physical abuse: A TAT analysis. Journal of Personality Assessment, 64, 552–568.

FREEDLE, R. (Ed.). (1990). Artificial intelligence and the future of testing. Hillsdale, NJ: Erlbaum.
FREILICH, M., RAYBECK, D., & SAVISHINSKY, J. (Eds.). (1991). Deviance: Anthropological perspectives, Westport, CT: Greenwood.
FRENCH, J. W. (1951). The description of aptitude and achievement tests in terms of rotated factors. Psychometric Monographs, No. 5.
FRENCH, J. W. (1962). Effect of anxiety on verbal and mathematical examination scores. Educational and Psychological Measurement, 22, 553–564.
FRENCH, J. W. (1965). The relationship of problem-solving styles to the factor composition of tests. Educational and Psychological Measurement, 25, 9–28.
FRENCH, J. W. (1966). The logic of and assumptions underlying differential testing. In A. Anastasi (Ed.), Testing problems in perspective (pp. 321–330). Washington, DC: American Council on Education.
FRISCH, M. B. (1994). QOLI—Quality of Life Inventory: Manual and treatment guide, Minneapolis, MN: National Computer Systems.
FRUZZETTI, A. E., &JACOBSON, N. S. (1992). Assessment of couples. In J. C. Rosen & R McReynolds (Eds.), Advances in psychological assessment (Vol. 8, pp. 201–224). New York: Plenum Press.
FRYER, D. (1931). Measurement of interests. New York: Holt.
FUCHS, L. S. (1993). Enhancing instructional programming and student achievement with curriculum-based measurement. In J. Kramer (Ed.), Curriculum-based measurement (pp. 65–104). Lincoln, NE: Buros Institute of Mental Measurements.
FUCHS, L. S., & DENO, S. L. (1991). Paradigmatic distinctions between instructionally relevant measurement models. Exceptional Children, 57, 488–500.
FUNDER, D. C. (1991). Explorations in behavioral consistency: Properties of persons, situations, and behaviors. Journal of Personality and Social Psychology, 60, 773–794.
FUNDER, D. C., PARKE, R. D., TOMLINSON-KEASEY, C., & WIDAMAN, K. (Eds.). (1993). Studying lives through time: Personality and development, Washington, DC: American Psychological Association.
FURLONG, M., & KARNO, M. (1995). Review of the Social Skills Rating System. Twelfth Mental Measurements Yearbook, 967–969.
FURNHAM, A. (1995). The relationship of personality and intelligence to cognitive learning style and achievement. In D. H. Saklofske & M. Zeidner (Eds.), International handbook of personality and intelligence (pp. 397–413). New York: Plenum Press.
GAGNE, R. (1965). The conditions of learning. New York: Holt, Rinehart & Winston.
GALTON, F. (1879). Psychometric experiments. Brain, 2, 149–162.
GALTON, F. (1883). Inquiries into human faculty and its development. London: Macmillan.
GAMBLE, K. R. (1972). The Holtzman Inkblot Technique: A review. Psychological Bulletin, 77, 172–194.
GARDNER, H. (1983). Frames of mind: The theory of multiple intelligences. New York: Basic Books.
GARDNER, H. (1992). Assessment in context: The alternative to standardized testing. In B. R. Gifford & M. C. O'Connor (Eds.), Changing assessments: Alternative views of aptitude, achievement and instruction (pp. 77–119). Boston: Kluwer.
GARDNER, H. (1993). Multiple intelligences: The theory in practice. New York: Basic Books.
GARDNER, J. W. (1961). Excellence. New York: Harper.
GATEWOOD, R. D., & FEILD, H. S. (1993). Human resource selection (3rd ed.). Chicago: Dryden Press.
GAUDRY, E., & SPIELBERGER, C. D. (1974). Anxiety and educational achievement. New York: Wiley.
GAUGLER, B. B., ROSENTHAL, D. B., THORNTON, G. C., III, & BENTSON, C. (1987). Meta-analysis of assessment center validity. Journal of Applied Psychology, 72, 493–511.
GDOWSKI, C. L., LACHAR, D., & KLINE, R. B. (1985). A PIG profile typology of children and adolescents: I. Empirically-derived alternative to traditional diagnosis. Journal of Abnormal Psychology, 94, 346–361.
GEARY, D. C. (1993). Mathematical disabilities: Cognitive, neuropsychological, and genetic components. Psychological Bulletin, 114, 345–362.
GEISINGER, K. F. (Ed.). (1992). Psychological testing of Hispanics. Washington, DC: American Psychological Association.
GEISINGER, K. F. (1994). Cross-cultural normative assessment: Translation and adaptation issues influencing the normative interpretation of assessment instruments. Psychological Assessment, 6, 304–312.
GELSO, C. J., & FRETZ, B. R. (1992). Counseling psychology. San Diego, CA: Harcourt Brace Jovanovich.
GENTILE, C. A., MARTIN-REHRMANN, J., & KENNEDY, J. H. (1995). Windows into the classroom: NAEP's 1992 writing portfolio study. Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement.
GERBER, M. M., SEMMEL, D. S., & SEMMEL, M. I. (1994). Computer-based dynamic assessment of multidigit multiplication. Exceptional Children, 61, 114–125.
GERGEN, K. J. (1985). The social constructionist movement in modern psychology. American Psychologist, 40, 266–275.
GESELL, A., et al. (1940). The first five years of life. New York: Harper.
GESELL, A., & AMATRUDA, C. S. (1947). Developmental diagnosis (2nd ed.). New York: Hoeber-Harper.
GHISELLI, E. E. (1956). Differentiation of individuals in terms of their predictability. Journal of Applied Psychology, 40, 374–377.
GHISELLI, E. E. (1959). The generalization of validity. Personnel Psychology, 12, 397–402.
GHISELLI, E. E. (1960). The prediction of predictability. Educational and Psychological Measurement, 20, 3–8.
GHISELLI, E. E. (1963). Moderating effects and differential reliability and validity. Journal of Applied Psychology, 47, 81–86.
GHISELLI, E. E. (1966). The validity of occupational aptitude tests. New York: Wiley.
GHISELLI, E. E. (1968). Interaction of traits and motivational factors in the determination of the success of managers. Journal of Applied Psychology, 52, 480–483.
GIFFORD, B. R. (Ed.). (1989a). Test policy and test performance: Education, language, and culture. Boston: Kluwer.
GIFFORD, B. R. (Ed.). (1989b). Test policy and the politics of opportunity allocation: The workplace and the law. Boston: Kluwer.
GIFFORD, B. R., & O'CONNOR, M. C. (Eds.). (1992). Changing assessments: Alternative views of aptitude, achievement, and instruction. Boston: Kluwer.

GILBERT, J. A. (1894). Researches on the mental and physical development of school children. Studies from the Yale Psychological Laboratory, 2, 40–100.
GINSBURG, H., & OPPER, S. (1969). Piaget's theory of intellectual development: An introduction. Englewood Cliffs, NJ: Prentice Hall.
GIRELLLI, S. A., & STAKE, J. E. (1993). Bipolarity in Jungian type theory and the Myers-Briggs Type Indicator. Journal of Personality Assessment, 60, 290–301.
GITOMER, D. H. (1993). Performance assessment and educational measurement. In R. E. Bennett & W. C. Ward (Eds.), Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment (pp. 241–263). Hillsdale, NJ: Erlbaum.
GLASER, R. (1963). Instructional technology and the measurement of learning outcomes. American Psychologist, 18, 519–522.
GLASER, R. (1984). Education and thinking: The role of knowledge. American Psychologist, 39, 93–104.
GLASS, G. V. (1976). Primary, secondary, and meta-analysis of research. Educational Researcher, 5, 3–8.
GLOBERSON, T., & ZELNIKER, T. (Eds.). (1989). Cognitive style and cognitive development. Norwood, NJ: Ablex.
GLUTTING, J. J., & KAPLAN, D. (1990). Stanford-Binet Intelligence Scale: Fourth Edi-tion: Making the case for reasonable interpretations. In C. R. Reynolds & R. W. Kamphaus (Eds.), Handbook of psychological and educational assessment of children: Intelligence and Achievement (pp. 277–295). New York: Guilford Press.
GLUTTING, J. J., & McDERMOTT, P. A. (1990). Principles and problems in learning potential. In C. R. Reynolds & R. W. Kamphaus (Eds.), Handbook of psychological and educational assessment of children: Intelligence and achievement (pp. 296–347). New York: Guilford Press.
GLUTTING, J. J., McDERMOTT, P. A., PRIFITERA, A., & McGRATH, E. A. (1994). Core profile types for the WISC-III and WIAT: Their development and application in identifying multivariate IQ-achievement discrepancies. School Psychology Review, 23, 619–639.
GLUTTING, J. J., McDERMOTT, P. A., PRIFITERA, A., & McGRATH, E. A. (1995). «Core profile types for the WISC-III and WIAT: Their development and application in identifying multivariate IQ-achievement discrepancies»: Errata. School Psychology Review, 24, 123–124.
GLUTTING, J. J., McDERMOTT, P. A., & STANLEY, J. C. (1987). Resolving differences among methods of establishing confidence limits for test scores. Educational and Psychological Measurement, 47, 607–614.
GLUTTING, J. J., & OAKLAND, T. (1992). Guide to the Assessment of Test Session Behavior for the WISC-III and the WIAT (GATSB), San Antonio, TX: Psychological Corporation.
GOETZ, E. T, & HALL, R. J. (1984). Evaluation of the Kaufman Assessment Battery for Children from an information-processing perspective. Journal of Special Education, 18, 281–296.
GOLDBERG, L. R. (1971). A historical survey of personality scales and inventories. In P. McReynolds (Ed.), Advances in psychological assessment (vol. 2, pp. 293–336). Palo Alto, CA: Science and Behavior Books.
GOLDBERG, L. R. (1991). Human mind versus regression equation: Five contrasts. In D. Cicchetti & W. M. Grove (Eds.), Thinking clearly about psychology: Essays in honor of Paul E. Meehl (Vol. 1, pp. 173–184). Minneapolis: University of Minnesota Press.
GOLDBERG, L. R. (1993). The structure of phenotypic personality traits. American Psychologist, 48, 26–34.
GOLDBERG, L. R., GRENIER, J. R., GUION, R. M., SECUREST, L. B., & WING, H. (1991). Questionnaires used in the prediction of trustworthiness in pre-employment selection decisions: An APA task force report, Washington, DC: American Psychological Association.
GOLDBERG, P. A. (1965). A review of sentence completion methods in personality assessment. Journal of Projective Techniques and Personality Assessment, 29, 12–45.
GOLDEN, C. J. (1981). The Luria-Nebraska Children's Battery: Theory and formulation. In G. W. Hynd & J. E. Obrzut (Eds.), Neuropsychological assessment and the school-age child: Issues and procedures (pp. 277–302). New York: Grune & Stratton.
GOLDEN, C. J. (1987). Computers in neuropsychology. In J. N. Butcher (Ed.), Computerized psychological assessment: A practitioner's guide (pp. 344–354). New York: Basic Books.
GOLDEN, C. J., PURISCH, A. D., & HAMMEKE, T. A. (1985). Luria-Nebraska Neuropsychological Battery: Forms I and II Manual, Los Angeles: Western Psychological Ser-vices.
GOLDEN, C. J., ZILLMER, E. A., & SPIERS, M. V. (1992). Neuropsychological assessment and intervention, Springfield, IL: Charles C Thomas.
GOLDFARB, R., & HALPERN, H. (1984). Word association responses in normal adult subjects. Journal of Psycholinguistic Research, 13, 37–55.
GOLDFRIED, M. R., & KENT, R. N. (1972). Traditional versus behavioral personality assessment: A comparison of methodological and theoretical assumptions. Psychological Bulletin, 77, 409–420.
GOLDING, S. L., & RORER, L. G. (1972). Illusory correlation and subjective judgment. Journal of Abnormal Psychology, 80, 249–260.
GOLDMAN, B. A., & MITCHELL, D. F. (1995). Directory of unpublished experimental mental measures (Vol. 6). Washington, DC: American Psychological Association.
GOLDSCHMID, M. L. (1968). Role of experience in the acquisition of conservation. Proceedings of the 76th Annual Convention of the American Psychological Association, 361–362.
GOLDSCHMID, M. L., & BENTLER, P. M. (1968a). Dimensions and measurements of conservation. Child Development, 39, 787–802.
GOLDSCHMID, M. L., & BENTLER, P. M. (1968b). Manual: Concept Assessment Kit — Conservation, San Diego, CA: Educational and Industrial Testing Service.
GOLDSCHMID, M. L., BENTLER, P. M., DEBUS, R. L., RAWLINSON, R., KOHNSTAMM, D., MODGIL, S., NICHOLLS, J. G., REYKOWSKI, J., STRUPCZEWSKA, B., WARREN, N. (1973). A cross-cultural investigation of conservation. Journal of Cross-Cultural Psychology, 4, 75–88.

GOLDSMITH, R. E., &NUGENT, N. (1984). Innovativeness and cognitive complexity: A second look. Psychological Reports, 55, 431-438.
GOLDSTEIN, R C., & LEVIN, H. S. (1985). Intellectual and academic outcome following closed head injury in children and adolescents: Research strategies and empirical findings. Developmental Neuropsychology, 1, 195-214.
GOLDSTEIN, G., & HERSEN, M. (Eds.) (1990). Handbook of psychological assessment (2nd ed.). New York: Pergamon Press.
GOLDSTEIN, I. L., ZEDECK, S., & SCHNEIDER, B. (1993). An exploration of the job analysis-content validity process. In N. Schmitt, W. C. Borman, et al. (Eds.), Personnel selection in organizations (pp. 3-34). San Francisco: Jossey-Bass.
GOLDSTEIN, K., & SCHEERER, M. (1941). Abstract and concrete behavior: An experimental study with special tests. Psychological Monographs, 53(2, Whole No. 230).
GOLDSTEIN, K. M., & BLACKMAN, S. (1978a). Assessment of cognitive style. In P. McReynolds (Ed.), Advances in psychological assessment (Vol. 4, pp. 462-525). San Francisco: Jossey-Bass.
GOLDSTEIN, K. M., & BLACKMAN, S. (1978b). Cognitive style: Five approaches and relevant research. New York: Wiley-Interscience.
GOLEMAN, D. (1995). Emotional intelligence. New York: Bantam Books.
GONCALVES, A. A., WOODWARD, M. J., & MILLON, T. (1994). Millon Clinical Multiaxial Inventory-II. In M. E. Maruish (Ed.), The use of psychological testing for treatment planning and outcome assessment (pp. 161-184). Hillsdale, NJ: Erlbaum.
GONZALEZ, R. (1996). Circles and squares, spheres and cubes: What's the deal with circumplex models? Journal of Vocational Behavior, 48, 77-84.
GOODENOUGH, D. R. (1976). The role of individual differences in field dependence as a factor in learning and memory. Psychological Bulletin, 83, 675-694.
GOODENOUGH, F. L. (1949). Mental testing: Its history, principles, and applications. New York: Rinehart.
GOODGLASS, H. (1986). The flexible battery in neuropsychological assessment. In T. Incagnoli, G. Goldstein, & C. J. Golden (Eds.), Clinical application of neuropsychological test batteries (pp. 121-134). New York: Plenum Press.
GOODMAN, J. R. (1990). Infant intelligence: Do we, can we, should we assess it? In C. R. Reynolds & R. W. Kamphaus (Eds.), Handbook of psychological and educational assessment of children: Intelligence and achievement (pp. 183-208). New York: Guilford Press.
GOODNOW, J. J. (1976). The nature of intelligent behavior: Questions raised by cross-cultural studies. In L. B. Resnick (Ed.), The nature of intelligence (pp. 169-188). Hillsdale, NJ: Erlbaum.
GOODYEAR, R. K. (1990). Research on the effects of test interpretation: A review. Counseling Psychologist, 18, 240-257.
GORDEN, R. L. (1992). Basic interviewing skills, Itasca, IL: F. E. Peacock.
GORDON, L. V., & ALF, E. F. (1960). Acclimatization and aptitude test performance. Educational and Psychological Measurement, 20, 333-337.
GORDON, M. A. (1953). A study of the applicability of the same minimum qualifying scores for technical schools to White males, WAF, and Negro males (Tech. Rep. No. 53-34). Lackland Air Force Base, TX: Personnel Research Laboratory.
GORMLY, A. V., & BRODZINSKY, D. M. (1993). Life-span human development (5th ed.). San Diego, CA: Harcourt Brace Jovanovich.
GOTTFREDSON, G. D. (1996). Prestige in vocational interests. Journal of Vocational Behavior, 48, 68-72.
GOTTFREDSON, G. D., & HOLLAND, J. L. (1989). Dictionary of Holland occupational codes (DHOC) (2nd ed.). Odessa, FL: Psychological Assessment Resources.
GOTTFREDSON, L. S. (Ed.). (1986a). The g factor in employment. Journal of Vocational Behavior, 29, 293-450.
GOTTFREDSON, L S. (1986b). Special groups and the beneficial use of vocational interest inventories. In W. B. Walsh & S. H. Osipow (Eds.), Advances in vocational psychology: Vol. 1. The assessment of interests (pp. 127-198). Hillsdale, NJ: Erlbaum.
GOTTFREDSON, L. S. (1994). The science and politics of race-norming. American Psychologist, 49, 955-963.
GOTTFRIED, A. W., & BRODY, N. (1975). Interrelationships between and correlates of psychometric and Piagetian scales of sensorimotor intelligence. Developmental Psychology, 11, 379-381.
GOTTMAN, J. M. (1994). What predicts divorce? The relationship between marital processes and marital outcomes, Hillsdale, NJ: Erlbaum.
GOTTMAN, J. M. (Ed.). (1995). The analysis of change, Hillsdale, NJ: Erlbaum.
GOTTMAN, J. M. (Ed.). (1996). What predicts divorce?: The measures, Mahwah, NJ: Erlbaum.
GOUGH, H. G. (1960). The Adjective Check List as a personality assessment research technique. Psychological Reports, 6, 107-122.
GOUGH, H. G. (1984). A managerial potential scale for the California Psychological Inventory. Journal of Applied Psychology, 69, 233-240.
GOUGH, H. G. (1985). A work orientation scale for the California Psychological Inventory. Journal of Applied Psychology, 70, 505-513.
GOUGH, H. G. (1987). California Psychological Inventory Administrator's guide, Palo Alto, CA: Consulting Psychologists Press.
GOUGH, H. G., & BRADLEY, P. (1996). CPI manual (3rd ed.). Palo Alto, CA: Consulting Psychologists Press.
GOUGH, H. G., & HEILBRUN, A. B., JR. (1980). The Adjective Check List bibliography, Palo Alto, CA: Consulting Psychologists Press.
GOUGH, H. G., & HEILBRUN, A. B., JR. (1983). The Adjective Check List manual (rev. ed.). Palo Alto, CA: Consulting Psychologists Press.
GRAHAM, J. R. (1993). MMPI-2: Assessing personality and psychopathology (2nd ed.). New York: Oxford University Press.
GRAVES, L. M. (1993). Sources of individual differences in interviewer effectiveness: A model and implications for future research. Journal of Organizational Behavior, 14, 349-370.
GRAVES, L M., & POWELL, G. N. (1988). An investigation of sex discrimination in recruiters evaluations of actual applications. Journal of Applied Psychology, 73, 20-29.

GRAYBILL, D. (1990). Developmental changes in the response types versus aggression categories on the Rosenzweig Picture-Frustration Study, Children's Form. Journal of Personality Assessment, 55, 603–609.
GRAYBILL, D. (1993). A longitudinal study of changes in children's thought content in response to frustration on the Children's Picture-Frustration Study. Journal of Personality Assessment, 61, 531–535.
GRE 1995–96 guide to the use of the Graduate Record Examinations Program, (1995). Princeton, NJ: Educational Testing Service.
GREDLER, G. R. (1992). School readiness: Assessment and educational issues. Brandon, VT: Clinical Psychology.
GREEN, B. F. (1983). The promise of tailored tests. In H. Wainer & S. Messick (Eds.), Principals of modern psychological measurement (pp. 69–80). Hillsdale, NJ: Erlbaum.
GREEN, B. E., BOCK, R. D., HUMPHREYS, L. G., LINN, R. L., & RECKASE, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. Journal of Educational Measurement, 22, 347–360.
GREEN, B. R, JR., & WIGDOR, A. K. (1991). Measuring job competency. In A. K. Wigdor & B. R Green, Jr. (Eds.), Performance assessment in the workplace: Vol. 2. Technical issues (pp. 53–74). Washington, DC: National Academy Press.
GREEN, D. R., FORD, M. P., & FLAMER, G. B. (Eds.) (1971). Measurement and Piaget: Proceedings of the CTB/McGraw-Hill Conference on Ordinal Scales of Cognitive Development. New York: McGraw-Hill.
GREENE, R. L. (1978). An empirically derived MMPI carelessness scale. Journal of Clinical Psychology, 34, 407–410.
GREENE, R. L. (1991). The MMPI-2/MMPI: An interpretive manual. Boston: Allyn & Bacon.
GREENO, J. G. (1989). A perspective on thinking. American Psychologist, 44, 134–141.
GREENWALD, G. (1982). Intelligence for peace: First international symposium on Venezuelan project for development of intelligence. Human Intelligence International Newsletter, 3 (6), pp. 1,3.
GREENWALD, G. (1984). Venezuelan ministry ends—Intelligence projects continue. Human Intelligence International Newsletter, 5(1), p. 1.
GREENWOOD, J. M., & McNAMARA, W. J. (1967). Interrater reliability in situational tests. Journal of Applied Psychology, 51, 101–106.
GREGG, N., HOY, C., & GAY, A. F. (Eds.). (1996). Adults with learning disabilities: Theoretical and practical perspectives. New York: Guilford Press.
GRESHAM, F. M., & ELLIOTT, S. N. (1990). Social Skills Rating System: Manual. Circle Pines, MN: American Guidance Service.
GRESHAM, F. M., ELLIOTT, S. N., & EVANS-FERNANDEZ, S. E. (1993). Student Self-Concept Scale: Manual. Circle Pines, MN: American Guidance Service.
GRESHAM, F. M., & LITTLE, S. G. (1993). Peer-referenced assessment strategies. In T H. Ollendick & M. Hersen (Eds.), Handbook of child and adolescent assessment (pp. 165–179). Boston: Allyn & Bacon.
GRESHAM, F M., MacMILLAN, D. L., & SIPERSTEIN, G. N. (1995). Critical analysis of the 1992 AAMR definition: Implications for school psychology. School Psychology Quarterly, 10, 1–19.
GRIBBONS, W. D., & LOHNES, P. R. (1982). Careers in theory and experience. Albany: State University of New York Press.
GRIGORENKO, E. L., & STERNBERG, R. J. (1995). Thinking styles. In D. H. Saklofske & M. Zeidner (Eds.), International handbook of personality and intelligence (pp. 205–229). New York: Plenum Press.
GROAT, L. (Ed.). (1995). Giving places meaning. San Diego, CA: Academic Press.
GROENVELD, M., & JAN, J. E. (1992). Intelligence profiles of low vision and blind children. Journal of Visual Impairment and Blindness, 86, 68–71.
GROOMS, R. R., & ENDLER, N. S. (1960). The effect of anxiety on academic achievement. Journal of Educational Psychology, 51, 299–304.
GROSS, A. L., FAGGEN, J., & McCARTHY, K. (1974). The differential predictability of the college performance of males and females. Educational and Psychological Measurement, 34, 363–365.
GROSS, A. L., & Su, W. H. (1975). Defining a «fair» or «unbiased» selection model: A question of utilities. Journal of Applied Psychology, 60, 345–351.
GROSSMAN, H. J. (Ed.). (1983). Classification in mental retardation. Washington, DC: American Association on Mental Retardation.
GROTH-MARNAT., G. (1990). Handbook of psychological assessment (2nd ed.). New York: Wiley.
GUERTIN, W. H., FRANK, G. H., & RABIN, A. I. (1956). Research with the Wechsler-Bellevue Intelligence Scale: 1950–1955. Psychological Bulletin, 53, 235–257.
GUERTIN, W. H., LADD, C. E., FRANK, G. H., RABIN, A. I., & HIESTER, D. S. (1966). Research with the Wechsler Intelligence Scale for Adults: 1960–1965. Psychological Bulletin, 66, 385–409.
GUERTIN, W. H., LADD, C. E., FRANK, G. H., RABIN, A. I., & HIESTER, D. S. (1971). Research with the Wechsler Intelligence Scale for Adults: 1965–1970. Psychological Record, 21, 289–339.
GUERTIN, W. H., RABIN, A. I., FRANK, G. H., & LADD, C. E. (1962). Research with the Wechsler Intelligence Scale for Adults: 1955–1960. Psychological Bulletin, 59, 1–26.
Guidelines for providers of psychological services to ethnic, linguistic, and culturally diverse populations. (1993). American Psychologist, 48, 45–48.
GUILFORD, J. P. (1959). Personality. New York: McGraw-Hill.
GUILFORD, J. P. (1967). The nature of human intelligence. New York: McGraw-Hill.
GUILFORD, J. P. (1981). Higher-order structure-of-intellect abilities. Multivariate Behavioral Research, 16, 411–435.
GUILFORD, J. P. (1988). Some changes in the Structure-of-Intellect Model. Educational and Psychological Measurement, 48, 1–4.
GUILFORD, J. P., & FRUCHTER, B. (1978). Fundamental statistics in psychology and education (6th ed.). New York: McGraw-Hill.
GUILFORD, J. P., & HOEPFNER, R. (1971). The analysis of intelligence. New York: McGraw-Hill.
GUILFORD, J. P. & ZIMMERMAN, W. S. (1956). Fourteen dimensions of temperament. Psychological Monographs, 70 (10, Whole No. 417).

GUION, R. M. (1991). Personnel assessment, selection, and placement. In M. D. Dunnette & L. M. Hough (Eds.), Handbook of industrial and organizational psychology (2nd ed., Vol. 2, pp. 327–397). Palo Alto, CA: Consulting Psychologists Press.
GUION, R. M., & GIBSON, W. M. (1988). Personnel selection and placement. Annual Review of Psychology, 39, 349–374.
GULLIKSEN, H. (1950). Theory of mental tests. New York: Wiley.
GULLIKSEN, H., & WILKS, S. S. (1950). Regression tests for several samples. Psychometrika,15, 91–114.
GUR, R. C., & GUR, R. E. (1991). The impact of neuroimaging on human neuropsychology. In R. G. Lister & H. J. Weingartner (Eds.), Perspectives in cognitive neuroscience (pp. 417–435). New York: Oxford University Press.
GUR, R. C., & GUR, R. E. (1994). Methods for the study of brain-behavior relationships. In A. Frazer, P. B. Molinoff, & A. Winokur (Eds.), Biological bases of brain function and disease (pp. 261–279). New York: Raven Press.
GUSTAFSSON, J.-E. (1984). A unifying model for the structure of intellectual abilities. Intelligence, 8, 179–203.
GUSTAFSSON, J.-E. (1989). Broad and narrow abilities in research on learning and instruction. In R. Kanfer, P. L. Ackerman, & R. Cudeck (Eds.), Abilities, motivation, and methodology (pp. 203–237). Hillsdale, NJ: Erlbaum.
GUTHRIE, G. M., JACKSON, D. N., ASTILLA, E., & ELWOOD, B. (1983). Personality measurement: Do the scales have similar meanings in another culture? In S. H. Irvine & J. W. Berry (Eds.), Human assessment and cultural factors (pp. 377–382). New York: Plenum Press.
GUTKIN, T. B., & REYNOLDS, C. R. (1981). Factorial similarity of the WISC-R for white and black children from the standardization sample. Journal of Educational Psychology, 73, 227–231.
GUTKIN, T. B., & WISE, S. (Eds.). (1991). The computer and the decision-making process. Hillsdale, NJ: Erlbaum.
GUTTMAN, I., & RAJU, N. S. (1965). A minimum loss function as determiner of optimal cutting scores. Personnel Psychology, 18, 179–185.
GUTTMAN, L. (1944). A basis for scaling qualitative data. American Sociological Review, 9,139–150.
GUTTMAN, L. (1947). The Cornell technique for scale and intensity analysis. Educational and Psychological Measurement, 7, 247–280.
GYURKE, J. S. (1991). The assessment of preschool children with the Wechsler Preschool and Primary Scale of Intelligence — Revised. In B. A. Bracken (Ed.), The psychoeducational assessment of preschool children (2nd ed., pp. 86–106). Boston: Allyn & Bacon.
HAAK, R. A. (1990). Using the sentence completion to assess emotional disturbance. In C. R. Reynolds & R. W. Kamphaus (Eds.), Handbook of psychological and educational assessment of children: Personality, behavior, and context (pp. 147–167). New York: Guilford Press.
HACKETT, G., & LONBORG, S. D. (1994). Career assessment and counseling for women. In W. B. Walsh & S. H. Osipow (Eds.), Career counseling for women (pp. 43–85). Hillsdale, NJ: Erlbaum.
HAFNER, J. L., FAKOURI, M. E., & LABRENTZ, H. L. (1982). First memories of «normal» and alcoholic individuals. Individual Psychology: Journal of Adlerian Theory, Research, and Practice, 38, 238–244.
HAGTVET, K. A., &JOHNSEN. T. B. (Eds.). (1992). Advances in test anxiety research (Vol. 7). Amsterdam: Swets & Zeitlinger.
HALADYNA, T. M. (1994). Developing and validating multiple-choice test items. Hillsdale, NJ: Erlbaum.
HALE, G. A., BRIDGEMAN, B., LEWIS, C., POLLACK, J. M., & WANG, M. (1992). A comparison of the predictive validity of the current SAT and an experimental prototype (ETS Res. Rep. 92–32). Princeton, NJ: Educational Testing Service.
HALSTEAD, W. C. (1947). Brain and intelligence. Chicago: University of Chicago Press.
HALVERSON, H. M. (1933). The acquisition of skill in infancy. Journal of Genetic Psychology, 43, 3–48.
HAMBLETON, R. K. (1984a). Determining test length. In R. A. Berk (Ed.), A guide to criterion-referenced test construction (pp. 144–168). Baltimore: Johns Hopkins University Press.
HAMBLETON, R. K. (1984b). Validating the test score. In R. A. Berk (Ed.), A guide to criterion-referenced test construction (pp. 199–230). Baltimore: Johns Hopkins University Press.
HAMBLETON, R. K. (1989). Principles and selected applications of item responses theory. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 147–200). New York: American Council on Education/Macmillan.
HAMBLETON, R. K. (1994). Guidelines for adapting educational and psychological tests: A progress report. European Journal of Psychological Assessment, 10, 229–244.
HAMBLETON, R. K. (1996). Guidelines for adapting tests (Final Report). Washington, DC: National Center for Education Statistics.
HAMBLETON, R. K., & NOVICK, M. R. (1973). Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement, 10, 159–170.
HAMBLETON, R. K., & ROGERS, H. J. (1989). Detecting potentially biased test items: Comparison of IRT area and the Mantel-Haenszel methods. Applied Measurement in Education, 2, 313–334.
HAMBLETON, R. K., SWAMINATHAN, H. S., & ROGERS, H. J. (1991). Fundamentals of item response theory. Newbury Park, CA: Sage.
HAMERS, J. H. M., SIJTSMA, K., & RUIJSSENAARS, A. J. J. M. (Eds.). (1993). Learning potential assessment: Theoretical, methodological, and practical issues. Amsterdam: Swets & Zeitlinger.
HAMILTON, J. L., & BUDOFF, M. (1974). Learning potential among the moderately and severely mentally retarded. Mental Retardation, 12, 33–36.
HAMILTON, R. G., & ROBERTSON, M. H. (1966). Examiner influence on the Holtzman Inkblot Technique. Journal of Projective Techniques and Personality Assessment, 30, 553–558.
HAMMER, E. R (1986). Graphic techniques with children and adolescents. In A. I. Rabin (Ed.), Projective techniques for adolescents and children (pp. 239–263). New York: Springer.
HANDLER, L. (1996). The clinical use of drawings: Draw-A-Person, House-Tree-Person, and Kinetic Family drawings. In C. S. Newmark (Ed.), Major psychological assessment instruments (2nd ed., pp. 206–293). Boston: Allyn & Bacon.
HANDLER, L., & HABENICHT, D. (1994). The Kinetic Family Drawing technique: A review of the literature. Journal of Personality Assessment, 62, 440–464.
HANDLER, L., & MEYER, G. J. (1996, Spring/Summer). Put your money where your mouth is! Mary Cerney's legacy. SPA Exchange, 6, 6–7.

HANNA, G. S., SONNENSCHEIN, J. L., & LENKE, J. M. (1983). The contribution of work-sample test items, student reported past grades, and student predicted grades in forecasting achievement in first-year algebra. Educational and Psychological Measurements, 43,243–249.
HANSEN, J. C. (1987). Cross-cultural research on vocational interests. Measurement and Evaluation in Counseling and Development, 19, 163–176.
HANSEN, J. C. (1990). Interest inventories. In G. Goldstein & M. Hersen (Eds.), Handbook of psychological assessment (2nd ed., pp. 173–194). New York: Pergamon Press.
HANSEN, J. C. (1996). What goes around, comes around. Journal of Vocational Behavior, 48, 73–76.
HANSEN, J. C., & CAMPBELL, D. P. (1985). Manual for the SVIB-SCII (4th ed.). Stanford, CA: Stanford University Press.
HANSON, F. A. (1993). Testing testing: Social consequences of the examined life. Berkeley: University of California Press.
HAPLIP, B., JR., & PANEK, P. E. (1993). Adult development and aging (2nd ed.). New York: Harper-Collins College.
HARDT, R. H., EYDE, L. D., PRIMOFF, E. S., & TORDY, G. R. (1981). The New York State Trooper job element examination: Final technical report. Albany: New York State Police. (National Technical Information Service, Springfield, VA 22161)
HARKNESS, A. R., McNULTY, J. L., & BEN-PORATH, Y. S. (1995). The Personality Psychopathology Five (PSY–5): Constructs and MMPI–2 scales. Psychological Assessment, 7,104–114.
HARLOW, H. F. (1949). The formation of learning sets. Psychological Review, 56, 51–65.
HARLOW, H. F. (1960). Learning set and error factor theory. In S. Koch (Ed.), Psychology: A study of a science (Vol. 2, pp. 492–537). New York: McGraw-Hill.
HARMAN, H. H. (1975). Final report of research on assessing human abilities (ONR Contract N00014–71-C-0117 Project NR 150 329). Princeton, NJ: Educational Testing Service.
HARMAN, H. H. (1976). Modern factor analysis (3rd ed.). Chicago: University of Chicago Press.
HARMON, L. W. (1996). Lost in space: A response to «The spherical representation of vocational interests» by Tracey and Rounds. Journal of Vocational Behavior, 48, 53–58.
HARMON, L. W., HANSEN, J. C., BORGEN, F. H., & HAMMER, A. L. (1994). Strong Interest Inventory: Applications and technical guide. Palo Alto, CA: Consulting Psychologists Press.
HARNQVIST, K. (1968). Relative changes in intelligence from 13 to 18. Scandinavian Journal of Psychology, 9, 50–82.
HARRE, R., & STEARNS, P. (Eds.) (1995). Discursive psychology in practice. Thousand Oaks, CA: Sage.
HARRINGTON, T. E, & O'SHEA, A. J. (1993). The Harrington-O'Shea Career Decision-Making System Revised: Manual. Circle Pines, MN: American Guidance Service.
HARRIS, D. B. (1963). Children's drawings as measures of intellectual maturity: A revision and extension of the Goodenough Draw-a-Man Test. San Diego, CA: Harcourt Brace Jovanovich.
HARRIS, J. A. (1973). The computer: Guidance tool of the future. In W. E. Coffman (Ed.), Frontiers of educational measurement and information systems—1973 (pp. 121–142). Boston: Houghton Mifflin.
HARRIS, M. J., & ROSENTHAL, R. (1985). Mediation of interpersonal expectancy effects: 31 meta-analyses. Psychological Bulletin, 97, 363–386.
HARRISON, P. L. (1985). Vineland Adaptive Behavior Scales: Classroom Edition manual. Circle Pines, MN: American Guidance Service.
HARRISON, R. (1965). Thematic apperception methods. In B. B. Wolman (Ed.), Handbook of clinical psychology (pp. 562–620). New York: McGraw-Hill.
HART, B., & RISLEY, T. R. (1995). Meaningful differences in the everyday experience of young American children. Baltimore: Brookes.
HART, D. H. (1986). The sentence completion techniques. In H. M. Knoff (Ed.), The assessment of child and adolescent personality (pp. 245–272). New York: Guilford Press.
HARTER, S. (1990). Issues in the assessment of the self-concept of children and adolescents. In A. M. La Greca (Ed.), Through the eyes of the child: Obtaining self-reports from children and adolescents (pp. 292–325). Boston: Allyn & Bacon.
HARTIGAN, J. A., & WIGDOR, A. K. (Eds.). (1989). Fairness in employment testing: Validity generalization minority issues, and the General Aptitude Test Battery. Washington, DC: National Academy Press.
HARTLE, T. W., & BATTAGLIA, P. A. (1993). The federal role in standardized testing. In R. E. Bennett & W. C. Ward (Eds.), Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment (pp. 291–311). Hillsdale, NJ: Erlbaum.
HARTMANN, D. P., & WOOD, D. D. (1990). Observational methods. In A. S. Bellack, M. Hersen, & A. E. Kazdin (Eds.), International handbook of behavior modification and therapy (2nd ed., pp. 107–138). New York: Plenum Press.
HARTSHORNE, H., & MAY, M. A. (1928). Studies in deceit. New York: Macmillan.
HARTSHORNE, H., MAY, M. A., & MALLER, J. B. (1929). Studies in service and self-control. New York: Macmillan.
HARTSHORNE, H., MAY, M. A., & SHUTTLEWORTH, F. K. (1930). Studies in the organization of character. New York: Macmillan.
HARVEY, R. J. (1991). Job analysis. In M. D. Dunnette & L. M. Hough (Eds.), Handbook of industrial and organizational psychology (2nd ed., Vol. 2, pp. 71–163). Palo Alto, CA: Consulting Psychologists Press.
HARVEY, R. J., & MURRY, W. D. (1994). Scoring the Myers-Briggs Type Indicator: Empirical comparison of preference score versus latent-trait methods. Journal of Personality Assessment, 62, 116–129.
HASKINS, R. (1989). Beyond metaphor: The efficacy of early childhood education. American Psychologist, 44, 274–282.
HASSELBLAD, V., & HEDGES, L. V. (1995). Meta-analysis of screening and diagnostic tests. Psychological Bulletin, 117, 167–178.
HATHAWAY, S. R., & McKINLEY, J. C. (1940). A Multiphasic Personality Schedule (Minnesota): I. Construction of the schedule. Journal of Psychology, 10, 249–254.
HATHAWAY, S. R., & MCKINLEY, J. C. (1943). The Minnesota Multiphasic Personality Inventory (rev. ed.). Minneapolis: University of Minnesota Press.
HATT, C. V. (1985). Review of Children's Apperception Test. Ninth Mental Measurements Yearbook, Vol. 1, 315–316.
HATTIE, J. (1992). Self-concept Hillsdale, NJ: Erlbaum.

HATTRUP, K. (1995). Review of the Differential Aptitude Tests, Fifth Edition. Twelfth Mental Measurements Yearbook, 302–304.
HAVILAND, J. (1976). Looking smart: The relationship between affect and intelligence in infancy. In M. Lewis (Ed.), Origins of intelligence: Infancy and early childhood (pp. 353–377). New York: Plenum Press.
HAWK, J. A. (1970). Linearity of criterion-GATB aptitude relationships. Measurement and Evaluation in Guidance, 2, 249–251.
HAYDUK, L. A. (1988). Structural equation modeling with LISREL: Essentials and advances. Baltimore: Johns Hopkins University Press.
HAYES, S. P. (1942). Alternative scales for the mental measurement of the visually handicapped. Outlook for the Blind, 36, 225–230.
HAYES, S. P. (1943). A second test scale for the mental measurement of the visually handicapped. Outlook for the Blind, 37, 37–41.
HAYNES, S. N. (1991). Behavioral assessment. In M. Hersen, A. E. Kazdin, & A. S. Bellack (Eds.), The clinical psychology handbook (2nd ed., pp. 430–464). New York: Pergamon Press.
HEATON, R. K., BAADE, L. E., & JOHNSON, K. L. (1978). Neuropsychological test results associated with psychiatric disorders in adults. Psychological Bulletin, 85, 141–162.
HEATON, R. K., GRANT, I., & MATTHEWS, C. G. (1991). Comprehensive norms for an expanded Halstead-Reitan battery. Odessa, FL: Psychological Assessment Resources.
HEBB, D. O. (1970). A return to Jensen and his social science critics. American Psychologist, 25, 568.
HEDGES, L. V. (1988). The meta-analysis of test validity studies: Some new approaches. In R. Wainer & H. I. Braun (Eds.), Test validity (pp. 191–212). Hillsdale, NJ: Erlbaum.
HEDGES, L. V. & NOWELL, A. (1995). Sex differences in mental test scores, variability, and numbers of high-scoring individuals. Science, 269, 41–45.
HEILBRUN, A. B., JR. (1964). Social-learning theory, social desirability, and the MMPI. Psychological Bulletin, 61, 377–387.
HEILBRUN, A. B., JR. (1985). Review of the California Child Q-Set. Ninth Mental Measurements Yearbook, Vol. 1, 248–249.
HEIN, M., & WESLEY, S. (1994). Scaling biodata through subgrouping. In G. S. Stokes, M. D. Mumford, & W. A. Owens (Eds.), Biodata handbook: Theory, research, and use of biographical information in selection and performance prediction (pp. 171–196). Palo Alto, CA: Consulting Psychologists Press.
HELFRICH, H. (1986). On linguistic variables influencing the understanding of questionnaire items. In A. Angleitner & J. S. Wiggins (Eds.), Personality assessment via questionnaires: Current issues in theory and measurement (pp. 178–188). Berlin: Springer-Verlag.
HELMES, E., & JACKSON, D. N. (1989). Prediction models of personality item responding. Multivariate Behavioral Research, 24, 71–91.
HELMES, E., & REDDON, J. R. (1993). A perspective on developments in assessing psychopathology: A critical review of the MMPI and MMPI-2. Psychological Bulletin, 113, 453–471.
HELSON, R., & WINK, P. (1992). Personality change in women from the early 40s to the early 50s. Psychology and Aging, 7, 46–55.
HENRY, B., MOFFITT, T. E., CASPI, A., LANGLEY, J., & SILVA, P. A. (1994). On the «Remembrance of things past»: A longitudinal evaluation of the retrospective method. Psychological Assessment, 6, 92–101.
HENRY, W. E. (1956). The analysis of fantasy: The thematic apperception technique in the study of personality. New York: Wiley.
HENRY, W. E., & PARLEY, J. (1959). The validity of the Thematic Apperception Test in the study of adolescent personality. Psychological Monographs, 73 (17, Whole No. 487).
HERMAN, S. J. (1994). Hiring right: A practical guide. Thousand Oaks, CA: Sage.
HERR, E. L. (1989). Review of the Kuder Occupational Interest Survey, Revised (Form DD). Tenth Mental Measurements Yearbook, 425–427.
HERRNSTEIN, R. J., & MURRAY, C. (1994). The bell curve: Intelligence and class structure in American life. New York: Free Press.
HERRNSTEIN, R. J., NICKERSON, R. S., SANCHEZ, M., & SWETS, J. A. (1986). Teaching thinking skills. American Psychologist, 41, 1279–1289.
HERRON, E. W. (1964). Changes in inkblot perception with presentation of the Holtzman inkblot technique as an «intelligence test.» Journal of Projective Techniques and Personality Assessment, 28, 442–447.
HERSEN, M., KAZDIN, A. E., & BELLACK, A. S. (Eds.). (1991). The clinical psychology handbook (2nd ed.). Elmsford, NY: Pergamon Press.
HERZBERGER, S. D., LINNEY, J. A., SEIDMAN, E., & RAPPAPORT, J. (1979). Preschool and primary locus of control scale: Is it ready for use? Developmental Psychology, 15, 320–324.
HESSEL, M. G. P., & HAMERS, J. H. M. (1993). The Learning Potential Test for Ethnic Minorities. In J. H. M. Hamers, K. Sijtsma, & A. J. J. M. Ruijssenaars (Eds.), Learning potential assessment: Theoretical, methodological, and practical issues. Amsterdam: Swets & Zeitlinger.
HETHERINGTON, E. M., REISS, D., & PLOMIN, R. (Eds.). (1993). The separate social worlds of siblings: The impact of nonshared environment on development. Hillsdale, NJ: Erlbaum.
HEWER, V. H. (1965). Are tests fair to college students from homes with low socioeconomic status? Personnel and Guidance Journal, 43, 764–769.
HIBBARD, S., FARMER, L., WELLS, C., DIFILLIPO, E., BARRY, W., KORMAN, R., & SLOAN, P. (1994). Validation of Cramer's defense mechanism manual for the TAT. Journal of Personality Assessment, 63, 197–210.
HICKS, L. E. (1970). Some properties of ipsative, normative, and forced normative measures. Psychological Bulletin, 74, 167–184.
HILL, D. J., & BALE, R. M. (1980). Development of the Mental Health Locus of Control and Mental Health Locus of Origin Scales. Journal of Personality Assessment, 44, 148–156.
HILL, E. F. (1972). Holtzman Inkblot Technique: A handbook for clinical application. San Francisco: Jossey-Bass.
HILL, K. T., & SARASON, S. B. (1966). The relation of test anxiety and defensiveness to test and school performance over the elementary school years. Monographs of the Society for Research in Child Development, 31, (2, Serial No. 104).
HILL, T. D., REDDON, J. R., & JACKSON, D. N. (1985). The factor structure of the Wechsler Scales: A brief review. Clinical Psychology Review, 5, 287–306.

HIRSH, S. K. (1995). Strong Interest Inventory resource: Strategies for group and individual interpretations in business and organizational settings. Palo Alto, CA: Consulting Psychologists Press.
HISKEY, M. S. (1966). The Hiskey Nebraska Test of Learning Aptitude. Lincoln, NE: Union College Press.
HOBBS, N. (1975a). The futures of children. San Francisco: Jossey-Bass.
HOBBS, N. (Ed.). (1975b). Issues in the classification of children (Vols. 1 & 2). San Francisco: Jossey-Bass.
HODAPP, R. M., BURACK, J. A., & ZIGLER, E. (Eds.). (1990). Issues in the developmental approach to mental retardation. New York: Cambridge University Press.
HODGES, K., & ZEMAN, J. (1993). Interviewing. In T. H. Ollendick & M. Hersen (Eds.), Handbook of child and adolescent assessment (pp. 65–81). Boston: Allyn & Bacon.
HOFER, P. J., & GREEN, B. F. (1985). The challenge of competence and creativity in computerized psychological testing. Journal of Consulting and Clinical Psychology, 53, 826–838.
HOFFMAN, B. (1962). The tyranny of testing. New York: Crowell-Collier.
HOFSTEE, W. K. B., DE RAAD, B., & GOLDBERG, L. R. (1992). Integration of the Big Five and circumplex approaches to trait structure. Journal of Personality and Social Psychology, 63,146–163.
HOGAN, J. C. (1992). Physical abilities. In M. D. Dunnette & L. M. Hough (Eds.), Handbook of industrial and organizational psychology (2nd ed., Vol. 2, pp. 753–831). Palo Alto, CA: Consulting Psychologists Press.
HOGAN, R. T. (1991). Personality and personality measurement. In M. D. Dunnette & L. M. Hough (Eds.), Handbook of industrial and organizational psychology (2nd ed., Vol. 2, pp. 873–919). Palo Alto, CA: Consulting Psychologists Press.
HOGAN, R., CURPHY, G. J., & HOGAN, J. (1994). What we know about leadership: Effectiveness and personality. American Psychologist, 49, 493–504.
HOGAN, R., DESOTO, S. B., & SOLANO, C. (1977). Traits, tests, and personality research. American Psychologist, 32, 255–264.
HOGAN, R., & HOGAN, J. (1992). Hogan Personality Inventory manual (2nd ed.). Tulsa, OK: Hogan Assessment Systems.
HOGAN, R., HOGAN, J., & ROBERTS, B. W. (1996). Personality measurement and employment decisions: Questions and answers. American Psychologist 51, 469–477.
HOGAN, R., & NICHOLSON, R. A. (1988). The meaning of personality test scores. American Psychologist, 43, 621–626.
HOLDEN, R. R., & JACKSON, D. N. (1992). Assessing psychopathology using the Basic Personality Inventory: Rationale and applications. In J. C. Rosen & R McReynolds (Eds.), Advances in psychological assessment (Vol. 8, pp. 165–199). New York: Plenum Press.
HOLLAND, J. L. (1959). A theory of vocational choice. Journal of Counseling Psychology, 59, 35–45.
HOLLAND, J. L. (1966). The psychology of vocational choice. Waltham, MA: Blaisdell.
HOLLAND, J. L. (1986). New directions for interest testing. In B. S. Plake & J. C. Witt (Eds.), The future of testing (pp. 245–267). Hillsdale, NJ: Erlbaum.
HOLLAND, J. L. (1992). Making vocational choices: A theory of vocational personalities and work environments (2nd ed.). Odessa, EL: Psychological Assessment Resources. (Original work published 1985)
HOLLAND, J. L., FRITZSCHE, B. A., & POWELL, A. B. (1994). The Self-Directed Search (SDS) Technical manual – 1994 edition. Odessa, EL: Psychological Assessment Resources.
HOLLAND, J. L., & GOTTFREDSON, G. D. (1976). Using a typology of persons and environments to explain careers; some extensions and clarification. Counseling Psychologist, 6, 20–29.
HOLLAND, J. L., & GOTTFREDSON, G. D. (1992). Studies of the hexagonal model: An evaluation (or, The perils of stalking the perfect hexagon). Journal of Vocational Behavior, 40, 158–170.
HOLLAND, J. L., POWELL, A. B., & FRITZSCHE, B. A. (1994). The Self-Directed Search (SDS) Professional user's guide–1994 edition, Odessa, FL: Psychological Assessment Resources.
HOLLAND, P. W., & RUBIN, D. B. (Eds.). (1982). Test equating. New York: Academic Press.
HOLLAND, P. W., & THAYER, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. Braun (Eds.), Test validity (pp. 129–145). Hillsdale, NJ: Erlbaum.
HOLLAND, P. W., & WAINER, H. (Eds.). (1993). Differential item functioning: Theory and Practice, Hillsdale, NJ: Erlbaum.
HOLLANDER, P. (1982). Legal context of educational testing. In A. K. Wigdor & W.R. Garner (Eds.), Ability testing: Uses, consequences, and controversies (pp. 195–231). Washington, DC: National Academic Press.
HOLLENBECK, G. P., & KAUFMAN, A. S. (1973). Factor analysis of the Wechsler Preschool and Primary Scale of Intelligence (WPPSI). Journal of Clinical Psychology, 29, 41–45.
HOLLENBECK, J. R., & WHITEMER, E. M. (1988). Criterion-related validation for small sample contexts: An integrated approach to synthetic validity. Journal of Applied Psychology, 73, 536–544.
HOLLINGSHEAD, A. B. (1957). Two-factor index of social position. Unpublished manuscript, Yale University, Department of Sociology, New Haven, CT.
HOLMSTROM, R. W., SILBER, D. E., & KARP, S. A. (1990). Development of the Apperceptive Personality Test. Journal of Personality Assessment, 54, 252–264.
HOLTZMAN, W. H. (1961). Guide to administration and scoring: Holtzman Inkblot Technique. New York: Psychological Corporation.
HOLTZMAN, W. H. (Ed.). (1970). Computer-assisted instruction, testing, and guidance. New York: Harper & Row.
HOLTZMAN, W. H. (1975). New developments in Holtzman Inkblot Technique. In P. McReynolds (Ed.), Advances in psychological assessment (Vol. 3, pp. 243–274). San Francisco: Jossey-Bass.
HOLTZMAN, W. H. (1986). Holtzman Inkblot Technique (HIT). In A. I. Rabin (Ed.), Assessment with projective techniques: A concise introduction, (pp. 47–83). New York: Springer.
HOLTZMAN, W. H. (1988). Beyond the Rorschach. Journal of Personality Assessment, 52, 578–609.
HOLTZMAN, W. H., MOSELEY, E. C., REINEHR, R. C., & ABBOTT, E. (1963). Comparison of the group method and the standard individual version of the Holtzman Inkblot Technique. Journal of Clinical Psychology, 19, 441–449.
HOLTZMAN, W. H., THORPE, J. S., SWARTZ, J. D., & HERRON, E. W. (1961). Inkblot perception and personality – Holtzman Inkblot Technique. Austin: University of Texas Press.

HONTS, C. R. (1994). Psychophysiological detection of deception. Current Directions in Psychological Science, 3, 77–82.
HONZIK, M. P. (1967). Environmental correlates of mental growth: Prediction from the family setting at 21 months. Child Development, 38, 337–364.
HONZIK, M. P., MACFARLANE, J. W., & ALLEN, L. (1948). The stability of mental test performance between two and eighteen years. Journal of Experimental Education, 17, 309–324.
HOOD, A. B., & JOHNSON, R. W. (1997). Assessment in counseling: A guide to the use of psychological assessment procedures (2nd ed.). Alexandria, VA: American Counseling Association.
HOOPER, F. H. (1973). Cognitive assessment across the life-span: Methodological implications of the organismic approach. In J. R. Nesselroade & H. W. Reese (Eds.), Life-span developmental psychology: Methodological issues (pp. 299–316). New York: Academic Press.
HOOPER, S. R., & WILLIS, W. C. (1989). Neuropsychological foundations, conceptual models, and issues in clinical differentiation. New York: Springer-Verlag.
HOPKINS, K. D., & STANLEY, J. C. (1981). Educational and psychological measurement and evaluation (6th ed.). Englewood Cliffs, NJ: Prentice Hall.
HORN, J. L. (1976). Human abilities: A review of research and theory in the early 1970s. Annual Review of Psychology, 27, 437–485.
HORN, J. L., & CATTELL, R. B. (1966). Refinement and test of the theory of fluid and crystallized general intelligences. Journal of Educational Psychology, 57, 253–270.
HORN, J. L., & KNAPP, J. R. (1973). On the subjective character of the empirical base of Guilford's structure-of-intellect model. Psychological Bulletin, 80, 33–43.
HORN, J. L., WANBERG, K. W., & FOSTER, F. M. (1990). Guide to the Alcohol Use Inventory (AU). Minneapolis, MN: National Computer Systems.
HOROWITZ, F D. (1994). The nature-nurture controversy in social and historical perspective. In F. Kessel (Ed.), Psychology, science, and human affairs Essays in honor of William Bevan (pp. 84–99). Boulder, CO: Westview Press.
HOROWITZ, F. D., & O'BRIEN, M. (Eds.). (1985). The gifted and talented: Developmental perspectives. Washington, DC: American Psychological Association.
HOROWITZ, F. D., & O'BRIEN, M. O. (Eds.). (1989). Children and their development: Knowledge base, research agenda, and social policy application [Special issue]. American Psychologist, 44 (2).
HORST, P. (1954). A technique for the development of a differential prediction battery. Psychological Monographs, 68 (9, Whole No. 380).
HOUGH, L. M., EATON, N. K., DUNNETTE, M. D., KAMP, J. D., & McCLOY, R. A. (1990). Criterion-related validities of personality constructs and the effect of response distortion on those validities. Journal of Applied Psychology Monograph, 75, 581–595.
HOUGH, L., & PAULLIN, C. (1994). Construct-oriented scale construction: The rational approach. In G. S. Stokes, M. D. Mumford, & W. A. Owens (Eds.), Biodata handbook: Theory, research, and use of biographical information in selection and performance prediction (pp. 109–145). Palo Alto, CA: Consulting Psychologists Press.
HOUSE, J. D. (1995). The predictive relationship between academic self-concept, achievement expectancies, and grade performance in college calculus. Journal of Social Psychology, 135, 111–112.
HOWARD, A., & BRAY, D. W. (1988). Managerial lives in transition: Advancing age and changing times. New York: Guilford Press.
HOWELL, D. C. (1997). Statistical methods for psychology (4th ed.). Belmont, CA: Duxbury Press.
HOWELL, K. W., & RUEDA, R. (1996). Achievement testing with culturally and linguistically diverse students. In L. A. Suzuki, P. J. Meller, & J. G. Ponterotto (Eds.), Handbook of multicultural assessment: Clinical, psychological, and educational applications (pp. 253–290). San Francisco: Jossey-Bass.
HRNCIR, E. J., SPELLER, G. M., & WEST, M. (1985). What are we testing? Developmental Psychology, 21, 226–232.
HU, S., & OAKLAND, T. (1991). Global and regional perspectives on testing children and youth: An empirical study. International Journal of Psychology, 26, 329–344.
HUGHES, J. (1990). Assessment of social skills: Sociometric and behavioral approaches. In C. R. Reynolds & R. W. Kamphaus (Eds.), Handbook of psychological and educational assessment of children: Personality, behavior, and context (pp. 423–444). New York: Guilford Press.
HULL, C. L. (1928). Aptitude testing. Yonkers, NY: World Book.
HUMPHREY, D. H., & DAHLSTROM, W. G. (1995). The impact of changing from the MMPI to the MMPI-2 on profile configurations. Journal of Personality Assessment, 64, 428–439.
HUMPHREYS, L. G. (1952). Individual differences. Annual Review of Psychology, 3, 131–150.
HUMPHREYS, L. G. (1962). The organization of human abilities. American Psychologist, 17, 475–483.
HUMPHREYS, L. G. (1970). A skeptical look at the factor pure test. In C. Lunneborg (Ed.), Current problems and techniques in multivariate psychology (pp. 23–32). Seattle: University of Washington Press.
HUMPHREYS, L. G. (1973). Statistical definitions of test validity for minority groups. Journal of Applied Psychology, 58, 1–4.
HUMPHREYS, L. G. (1979). The construct of general intelligence. Intelligence, 3, 105–120.
HUMPHREYS, L. G., RICH, S. A., & DAVEY, T. C. (1985). A Piagetian test of general intelligence. Developmental Psychology, 21, 872–877.
HUNT, E. (1987). Science, technology, and intelligence. In R. R. Ronning, J. A. Glover, J. C. Conoley, & J. C. Witt (Eds.), The influence of cognitive psychology on testing (pp. 11–40). Hillsdale, NJ: Erlbaum.
HUNT, J. McV. (1976). The utility of ordinal scales inspired by Piaget's observations. Merrill-Palmer Quarterly, 22, 31–45.
HUNT, J. McV. (1981). Experiential roots of intention, initiative, and trust. In H. I. Day (Ed.), Advances in intrinsic motivation and aesthetics (pp. 169–202). New York: Plenum Press.
HUNT, J. McV., & KIRK, G. E. (1974). Criterion-referenced tests of school readiness: A paradigm with illustrations. Genetic Psychology Monographs, 90, 143–182.
HUNTER, J. E. (1986). Cognitive ability, cognitive aptitudes, job knowledge, and job performance. Journal of Vocational Behavior, 29, 340–362.

HUNTER, J. E., & HUNTER, R. F. (1984). Validity and utility of alternative predictors of job performance. Psychological Bulletin, 96, 72–98.
HUNTER, J. E., & SCHMIDT, F. L. (1976). Critical analysis of the statistical and ethical implications of various definitions of test bias. Psychological Bulletin, 83, 1053–1071.
HUNTER, J. E., & SCHMIDT, F. L. (1981). Fitting people into jobs: The impact of personnel selection on national productivity. In M. A. Dunnette & E. A. Fleishman (Eds.), Human performance and productivity: Vol. 1. Human capability assessment (pp. 233–284). Hillsdale, NJ: Erlbaum.
HUNTER, J. E., & SCHMIDT, F. L. (1990). Methods of meta-analysis: Correcting error and bias in research findings. Newbury Park, CA: Sage.
HUNTER, J. E., SCHMIDT, F. L., & HUNTER, R. (1979). Differential validity of employment tests by race: A comprehensive review and analysis. Psychological Bulletin, 86, 721–735.
HUNTER, J. E., SCHMIDT, F. L., & JUDIESCH, M. K. (1990). Individual differences in output variability as a function of job complexity. Journal of Applied Psychology, 75, 28–42.
HUNTER, J. E., SCHMIDT, F. L., & RAUSCHENBERGER, J. M. (1977). Fairness of psychological tests: Implications of four definitions for selection utility and minority hiring. Journal of Applied Psychology, 62, 245–260.
HUNTER, J. E., SCHMIDT, F. L., & RAUSCHENBERGER, J. (1984). Methodological, statistical, and ethical issues in the study of bias in psychological tests. In C. E. Reynolds & R. T Brown (Eds.), Perspectives on bias in mental testing (pp. 41–99). New York: Plenum Press.
HURT, S. W., REZNIKOFF, M., & CLARKIN, J. F. (1991). Psychological assessment, psychiatric diagnosis, and treatment planning. New York: Brunner-Mazel.
HUSEN, T. (1951). The influence of schooling upon IQ. Theoria, 17, 61–88.
HY, L. X., & LOEVINGER, J. (1996). Measuring ego development (2nd ed.). Mahwah, NJ: Erlbaum.
INHELDER, B., DE CAPRONA, D., & CORNU-WELLS, A. (Eds.). (1987). Piaget today. Hove, England: Erlbaum.
INTELLIGENCE AND ITS MEASUREMENT: A SYMPOSIUM. (1921). Journal of Educational Psychology, 12, 123–147, 195–216.
IRETON, H., THWING, E., & GRAVEM, H. (1970). Infant mental development and neurological status, family socioeconomic status, and intelligence at age four. Child Development, 41, 937–945.
IRONSON, G. H., GUION, R. M., & OSTRANDER, M. (1982). Adverse impact from a psychometric perspective. Journal of Applied Psychology, 67, 419–432.
IRVINE, S. H. (1969a). Factor analyses of African abilities and attainments: Constructs across cultures. Psychological Bulletin, 71, 20–32.
IRVINE, S. H. (1969b). Figural tests of reasoning in Africa: Studies in the use of Raven's matrices across cultures. International Journal of Psychology, 4, 217–228.
IRVINE, S. H. (1983). Testing in Africa and America. In S. H. Irvine & J. W. Berry (Eds.), Human assessment and cultural factors (pp. 45–58). New York: Plenum Press.
IRVINE, S. H., & BERRY, J. W. (Eds.). (1988). Human abilities in cultural contexts. New York: Cambridge University Press.
IRVINE, S. H., & CARROLL, W. K. (1980). Testing and assessment among cultures: Issues in methodology and theory. In H. C. Triandis et al. (Eds.), Handbook of cross-cultural psychology (Vol. 2, pp. 181–244). Boston: Allyn & Bacon.
ISAACS, M., & CHEN, K. (1990). Presence/absence of an observer in a word association test. Journal of Personality Assessment, 55, 41–51.
IVNIK, R. J., MALEC, J. E, SMITH, G. E., TANGALOS, E. G., PETERSEN, R. C., KORMEN, E., & KURLAND, L. T. (1992). Mayo's older Americans normative studies: WAIS-R norms for ages 56 to 97. Clinical Neuropsychologist, 6 (Suppl.), 1–30.
IZARD, C. E., KAGAN, J., & ZAJONC, R. B. (Eds.). (1989). Emotions, cognition, and behavior. New York: Cambridge University Press.
JACKSON, D. N. (1970). A sequential system for personality scale development. In C. D. Spielberger (Ed.), Current topics in clinical and community psychology (Vol. 2, pp. 61–96). New York: Academic Press.
JACKSON, D. N. (1971). The dynamics of structured personality tests. Psychological Review, 78, 229–248.
JACKSON, D. N. (1973). Structured personality assessment. In B. B. Wolman (Ed.), Handbook of general psychology (pp. 775–792). Englewood Cliffs, NJ: Prentice Hall.
JACKSON, D. N. (1976) Jackson Personality Inventory: Manual. Port Huron, MI: Research Psychologists Press.
JACKSON, D. N. (1977). Jackson Vocational Interest Survey manual. Port Huron, MI: Research Psychologists Press.
JACKSON, D. N. (1985). Computer-based personality testing. Computers in Human Behavior, 1, 225–264.
JACKSON, D. N. (1986a). Career Directions Inventory manual. Port Huron, MI: Research Psychologists Press.
JACKSON, D. N. (1986b). The process of responding in personality assessment. In A. Angleitner & J. S. Wiggins (Eds.), Personality assessment via questionnaires: Current issues in theory and measurement (pp. 123–142). Berlin: Springer-Verlag.
JACKSON, D. N. (1989a). Basic Personality Inventory: BPI manual. Port Huron, MI: Sigma Assessment Systems.
JACKSON, D. N. (1989b). Personalit:y Research Form manual (3rd ed.). Port Huron, MI: Sigma Assessment Systems.
JACKSON, D. N. (1991). Computer-assisted personality test interpretation: The dawn of discovery. In T B. Gutkin & S. L. Wise (Eds.), The computer and the decision-making process (pp. 1–10). Hillsdale, NJ: Erlbaum.
JACKSON, D. N. (1994a). Jackson Personality Inventory-Revised: Manual. Port Huron, MI: Sigma Assessment Systems.
JACKSON, D. N. (1994b). Multidimensional Aptitude Battery (MAB): Manual. Port Huron, MI.: Sigma Assessment Systems. (1st ed., 1984)
JACKSON, D. N. (1995). JVIS occupations guide. Port Huron, MI: Sigma Assessment Systems.
JACKSON, D. N., GUTHRIE, G. M., ASTILLA, E., & ELWOOD, B. (1983). The cross-cultural generalization of personality construct measures. In S. H. Irvine & J. W. Berry (Eds.), Human assessment and cultural factors (pp. 365–375). New York: Plenum Press.
JACKSON, D. N., & MESSICK, S. (1958). Content and style in personality assessment. Psychological Bulletin, 55, 243–252.
JACKSON, D. N., & MESSICK, S. (1962). Response styles and the assessment of psychopathology. In S. Messick & J. Ross (Eds.), Measurement in personality and cognition (pp. 129–155). New York: Wiley.

JACKSON, D. N., & PAUNONEN, S. V. (1980). Personality structure and assessment. Annual Review of Psychology, 31, 503–551.
JACKSON, D. N., & WILLIAMS, D. R. (1975). Occupational classification in terms of interest patterns. Journal of Vocational Behavior, 6, 269–280.
JACOB, S., & HARTSHORNE, T. S. (1991). Ethics and law for school psychologists. Brandon, VT: Clinical Psychology Publishing Co.
JACOBS, A., & BARRON, R. (1968). Falsification of the Guilford-Zimmerman Temperament Survey: II. Making a poor impression. Psychological Reports, 23, 1271–1277.
JACOBS, P. I., & VANDEVENTER, M. (1971). The learning and transfer of double-classification skills: A replication and extension. Journal of Experimental Child Psychology, 12,140–157.
JACOBSON, J. W., & MULICK, J. A. (Eds.). (1996). Manual of diagnosis and professional practice in mental retardation. Washington, DC: American Psychological Association.
JAEGER, R. M. (1973). The national test-equating study in reading (The Anchor Test Study). NCME Measurement in Education, 4 (4), 1–8.
JAEGER, R. M. (Ed.) (1977). Applications of latent trait models [Special issue]. Journal of Educational Measurement, 14 (2).
JAEGER, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 485–514). New York: American Council on Education/Macmillan.
JAMES, L. A., & JAMES, L. R. (1989). Integrating work environment perceptions: Explorations into the measurement of meaning. Journal of Applied Psychology, 74, 739–751.
JAMES, L. R. (1973). Criterion models and construct validity for criteria. Psychological Bulletin, 80, 75–83.
JAMES, L. R. (1980). The unmeasured variable problem in path analysis. Journal of Applied Psychology, 65, 415–421.
JAMES, L. R., DEMAREE, R. G., MULAIK, S. A., & LADD, R. T. (1992). Validity generalization in the context of situational models. Journal of Applied Psychology, 77, 3–14.
JAMES, L. R., MULAIK, S. A., & BRETT, J. M. (1982). Causal analysis: Assumptions, models, and data. Beverly Hills, CA: Sage Publications.
JAYNES, J. H., & WLODKOWSKI, R. J. (1990). Eager to learn: Helping children become motivated and love learning. San Francisco: Jossey-Bass.
JENKINS, J. J., & RUSSELL, W. A. (1960). Systematic changes in word association norms: 1910–1952. Journal of Abnormal Psychology, 60, 293–304.
JENSEN, A. R. (1968). Social class and verbal learning. In M. Deutsch, I. Katz, & A. R. Jensen (Eds.), Social class, race, and psychological development (pp. 115–174). New York: Holt, Rinehart & Winston.
JENSEN, A. R. (1969). How much can we boost IQ and scholastic achievement? Harvard Educational Review, 39, 1–123.
JENSEN, A. R. (1984). The black-white difference on the K-ABC: Implications for future tests. Journal of Special Education, 18, 377–408.
JITENDRA, A. K., KAMEENUI, E. J., & CARNINE, D. W. (1994). An exploratory evaluation of dynamic assessment and the role of basals on comprehension of mathematical operations. Education and Treatment of Children, 17, 139–162.
JOHANSSON, C. B. (1984). Career Assessment Inventory: The Vocational Version (2nd ed.). Minneapolis, MN: National Computer Systems.
JOHANSSON, C. B. (1986). Career Assessment Inventory: The Enhanced Version. Minneapolis, MN: National Computer Systems.
JOHN, O. P., ANGLEITNER, A., & OSTENDORF, F. (1988). The lexical approach to personality: A historical review of trait taxonomic research. European Journal of Personality, 2,171–203.
JOHNSON, A. P. (1951). Notes on a suggested index of item validity: The U-L index. Journal of Educational Psychology, 42, 499–504.
JOHNSON, D. L., SWANK, P., HOWIE, V. M., BALDWIN, C. D., OWEN, M., & LUTTMAN, D. (1993). Does the HOME add to the prediction of child intelligence over and above SES? Journal of Genetic Psychology, 154, 33–40.
JOHNSON, E. G. (1992). The design of the National Assessment of Educational Progress. Journal of Educational Measurement, 29, 95–110.
JOHNSON, N. L, & GOLD, S. N. (1995). The Defense Mechanism Profile: A sentence completion test. In H. R. Conte & R. Plutchik (Eds.), Ego defenses: Theory and measurement (pp. 247–262). New York: Wiley.
JOINT COMMITTEE ON TESTING PRACTICES (JCTP). (1988). Code of fair testing practices in education. Washington, DC: Author. (Information about the Joint Committee is available from the Joint Committee on Testing Practices, American Psychological Association, 750 First Street, NE, Washington, DC 20002.)
JONASSEN, D. H., & GRABOWSKI, B. L. (1993). Handbook of individual differences, learning, and instruction. Hillsdale, NJ: Erlbaum.
JONES, L. E., & KOEHLY, L. M. (1993). Multidimensional scaling. In G. Keren & C. Lewis (Eds.), A handbook for data analysis in the behavioral sciences: Methodological issues (pp. 95–163). Hillsdale, NJ: Erlbaum.
JONES, L. V., & APPELBAUM, M. I. (1989). Psychometric methods. Annual Review of Psychology, 40, 23–43.
JONES, P. B., & SABERS, D. L. (1992). Examining test data using multivariate procedures. In M. Zeidner & R. Most (Eds.), Psychological testing: An inside view (pp. 297–339). Palo Alto, CA: Consulting Psychologists Press.
JONES, R. R., REID, J. B., & PATTERSON, G. R. (1975). Naturalistic observation in clinical assessment. In P. McReynolds (Ed.), Advances in psychological assessment (Vol. 3, pp. 42–95). San Francisco: Jossey-Bass.
JORESKOG, K. G., & SORBOM, D. (1986). LISREL: Analysis of linear structural relationships by maximum likelihood, instrumental variables, and least squares methods (4th ed.). Mooresville, IN: Scientific Software.
JORESKOG, K. G., & SORBOM, D. (1989). LISREL 7 Users Guide. Mooresville, IN: Scientific Software.
JORESKOG, K. G., & SЦRBOM, D. (1993). LISREL 8 structural equation modeling with the SIMPLIS command language. Hillsdale, NJ: Erlbaum.
JUNG, C. G. (1910). The association method. American Journal of Psychology, 21, 219–269.

JUNG, C. G. (1971). Psychological types (H. G. Baynes, Trans. revised by R. F. C. Hull). Princeton, NJ: Princeton University Press. (Original work published 1921)

KAGAN, J. (1965). Impulsive and reflective children: Significance of conceptual tempo. In J. Krumboltz (Ed.), Learning and the educational process (pp. 133–161). Chicago: Rand McNally.

KAGAN, J., & FREEMAN, M. (1963). Relation of childhood intelligence, maternal behaviors, and social class to behavior during adolescence. Child Development, 34, 899–911.

KAGAN, J., SONTAG, L. W., BAKER, C. T., & NELSON, V. L. (1958). Personality and IQ change. Journal of Abnormal and Social Psychology, 56, 261–266.

KAHN, J. V. (1987). Uses of the scales with mentally retarded populations. In I. C. Užgiris & J. McV. Hunt (Eds.), Infant performance and experience: New findings with the ordinal scales (pp. 252–280). Champaign: University of Illinois Press.

KAISER, H. F. (1958). A modified stanine scale. Journal of Experimental Education, 26, 261.

KAISER, H. F., & MICHAEL, W. B. (1975). Domain validity and generalizability. Educational and Psychological Measurement, 35, 31–35.

KAMINER, Y., FEINSTEIN, C., & SEIFER, R. (1995). Is there a need for observationally based assessment of affective symptomatology in child and adolescent psychiatry? Adolescence, 30, 483–489.

KAMPHAUS, R. W. (1990). K-ABC theory in historical and current contexts. Journal of Psychoeducational Assessment, 8, 356–368.

KAMPHAUS, R. W. (1993). Clinical assessment of children's intelligence: A handbook for professional practice. Boston: Allyn & Bacon.

KAMPHAUS, R. W., & FRICK, P. J. (1996). Clinical assessment of child and adolescent personality and behavior. Boston: Allyn & Bacon.

KAMPHAUS, R. W., KAUFMAN, A. S., & HARRISON, P. L. (1990). Clinical assessment practice with the Kaufman Assessment Battery for Children (K-ABC). In C. R. Reynolds & R. W. Kamphaus (Eds.), Handbook of psychological and educational assessment of children: Intelligence and achievement, (pp. 259–276). New York: Guilford Press.

KAMPHAUS, R. W., & REYNOLDS, C. R. (1987). Clinical and research applications of the K-ABC. Circle Pines, MN: American Guidance Service.

KANE, J. S., & LAWLER, E. E., III. (1978). Methods of peer assessment. Psychological Bulletin, 85, 555–586.

KANFER, R., ACKERMAN, P. L., & CUDECK, R. (Eds.). (1989). Abilities, motivation, and methodology (The Minnesota Symposium on Learning and Individual Differences). Hillsdale, NJ: Erlbaum.

KANFER, R., ACKERMAN, P. L., MURTHA, T, & GOFF, M. (1995). Personality and intelligence in industrial and organizational psychology. In D. H. Saklofske & M. Zeidner (Eds.), International handbook of personality and intelligence (pp. 577–602). New York: Plenum Press.

KANTOR, J. E., & CARRETTA, T. R. (1988). Aircrew selection systems. Aviation, Space, and Environmental Medicine, 59, 32–38.

KAPES, J. T., MASTIE, M. M., & WHITFIELD, E. A. (Eds.). (1994). A counselor's guide to career assessment instruments (3rd ed.). Alexandria, VA: National Career Development Association.

KAPLAN, M. E, & ERON, L. D. (1965). Test sophistication and faking in the TAT situation. Journal of Projective Techniques, 29, 498–503.

KARLSEN, B. (1992). LAAP, Language Arts Assessment Portfolio: Teachers guide (Levels I-III). Circle Pines, MN: American Guidance Service.

KARNES, F. A., & BROWN, K. E. (1980). Factor analysis of WISC-R for the gifted. Journal of Educational Psychology, 72, 197–199.

KATZ, M. R. (1974). Career decision-making: A computer-based System of Interactive Guidance and Information (SIGI). Proceedings of the 1973 Invitational Conference on Testing Problems, Educational Testing Service, 43–69.

KATZ, M. R. (1993). Computer-assisted career decision making: The guide in the machine. Hillsdale, NJ: Erlbaum.

KATZ, S., & LAUTENSCHLAGER, G. J. (1995). The SAT reading task in question: Reply to Freedle and Kostin. Psychological Science, 6, 126–127.

KAUFMAN, A. S. (1971). Piaget and Gesell: A psychometric analysis of tests built from their tasks. Child Development, 42, 1341–1360.

KAUFMAN, A. S. (1975). Factor analysis of the WISC-R at eleven age levels between 6 1/2 and 16 1/2 years. Journal of Counseling and Clinical Psychology, 43, 135–147.

KAUFMAN, A. S. (1979). Intelligent testing with the WISC-R. New York: Wiley.

KAUFMAN, A. S. (1990). Assessing adolescent and adult intelligence. Boston: Allyn & Bacon.

KAUFMAN, A. S. (1994). Intelligent testing with the WISCT-III. New York: Wiley.

KAUFMAN, A. S., & HOLLENBECK, G. P. (1974). Comparative structure of the WPPSI for blacks and whites. Journal of Clinical Psychology, 30, 316–319.

KAUFMAN, A. S., & KAUFMAN, N. L. (1972). Tests built from Piaget's and Gesell's tasks as predictors of first-grade achievement. Child Development, 43, 521–535.

KAUFMAN, A. S., & KAUFMAN, N. L. (1977). Clinical evaluation of young children with the McCarthy Scales. New York: Grune & Stratton.

KAUFMAN, A. S., & KAUFMAN, N. L. (1983a). Kaufman Assessment Battery for Children: Administration and scoring manual. Circle Pines, MN: American Guidance Service.

KAUFMAN, A. S., & KAUFMAN, N. L. (1983b). Kaufman Assessment Battery for Children: Interpretive manual. Circle Pines, MN: American Guidance Service.

KAUFMAN, A. S., & KAUFMAN, N. L. (1985). Kaufman Test of Educational Achievement: Comprehensive Form manual. Circle Pines, MN: American Guidance Service.

KAUFMAN, A. S., & KAUFMAN, N. L. (1990). Kaufman Brief Intelligence Test: Manual. Circle Pines, MN: American Guidance Service.

KAUFMAN, A. S., & KAUFMAN, N. L. (1993). Kaufman Adolescent and Adult Intelligence Test: Manual. Circle Pines, MN: American Guidance Service.

KAUSLER, D. H. (1994). Learning and memory in normal aging. San Diego, CA: Academic Press.
KAVALE, K. A., & FORNESS, S. R. (1984). A meta-analysis of the validity of Wechsler Scale profiles and recategorizations: Patterns or parodies? Learning Disability Quarterly, 7, 136–156.
KAVRUCK, S. (1956). Thirty-three years of test research: A short history of test development in the U. S. Civil Service Commission. American Psychologist, 11, 329–333.
KEHOE, J. F. (1992). Review of the Career Assessment Inventory, Second Edition [Vocational version]. Eleventh Mental Measurements Yearbook, 149.
KEHOE, J. F., & TENOPYR, M. L. (1994). Adjustment in assessment scores and their usage: A taxonomy and evaluation of methods. Psychological Assessment, 6, 291–303.
KEISER, R. E., & PRATHER, E. N. (1990). What is the TAT? A review often years of research. Journal of Personality Assessment, 55, 800–803.
KEITH, T. Z. (1985). Questioning the K-ABC: What does it measure? School Psychology Review, 14, 9–20.
KEITH, T. Z., & DUNBAR, S. B. (1984). Hierarchical factor analysis of the K-ABC: Testing alternate models. Journal of Special Education, 18, 367–375.
KELLER, L. S., & BUTCHER, J. N. (1991). Assessment of chronic pain patients with the MMPI–2. Minneapolis: University of Minnesota Press.
KELLEY, C., & MEYERS, J. E. (1993). The Cross-Cultural Adaptability Inventory. Minneapolis, MN: National Computer Systems.
KELLEY, M. R, & SURBECK, E. (1991). History of preschool assessment. In B. A. Bracken (Ed.), The psychoeducational assessment of preschool children (2nd ed., pp. 1–17). Boston: Allyn & Bacon.
KELLEY, T. L. (1928). Crossroads in the mind of man: A study of differentiable mental abilities. Stanford, CA: Stanford University Press.
KELLEY, T L. (1935). Essential traits of mental life. Cambridge, MA: Harvard University Press.
KELLEY, T. L (1939). The selection of upper and lower groups for the validation of test items. Journal of Educational Psychology, 30, 17–24.
KELLEY, T. L (1943). Cumulative significance of a number of independent experiments: Reply to A. E. Traxler and R. N. Hilkert. School and Society, 57, 482–484.
KELLY, G. A. (1955). The psychology of personal constructs. New York: Norton.
KELLY, G. A. (1963). A theory of personality. New York: Norton.
KELLY, G. A. (1970). A summary statement of a cognitively oriented comprehensive theory of behavior. In J. C. Mancuso (Ed.), Readings for a cognitive theory of personality (pp. 27–58). New York: Holt, Rinehart & Winston.
KELLY, M. P., & MELTON, G. B. (1993). Legal and ethical issues: In J. L. Culbertson & D. J. Willis (Eds.), Testing young children: A reference guide for developmental, psychoeducational, and psychosocial assessments (pp. 408–425). Austin, TX: PRO-ED.
KELZ, J. W. (1966). The development and evaluation of a measure of counselor effectiveness. Personnel and Guidance Journal, 44, 511–516.
KENRICK, D. T., & FUNDER, D. C. (1988). Profiting from controversy: Lessons from the person-situation debate. American Psychologist, 43, 23–34.
KENT, G. H., & ROSANOFF, A. J. (1910). A study of association in insanity. American Journal of Insanity, 67, 37–96, 317–390.
KENT, R. N., & FOSTER, S. L. (1977). Direct observational procedures: Methodological issues in naturalistic settings. In A. R. Ciminero, K. S. Calhoun, & H. E. Adams (Eds.), Handbook of behavioral assessment (pp. 279–328). New York: Wiley.
KERLINGER, F. N. (1986). Foundations of behavioral research (3rd ed.). New York: Holt, Rinehart & Winston.
KEYSER, D. J., & SWEETLAND, R. C. (Eds.). (1984–1994). Test critiques. Austin, TX: PRO-ED.
KHAN, S. B. (1970). Development of mental abilities: An investigation of the «differentiation hypothesis.» Canadian Journal of Psychology, 24, 199–205.
KHAN, S. B. (1972). Learning and the development of verbal ability. American Educational Research Journal, 9, 607–614.
KHAN, S. B., ALVI, S. A., SHAUKAT, N., & HUSSAIN, M. A. (1990). A study of the validity of Holland's theory in a non-Western culture. Journal of Vocational Behavior, 36, 132–146.
KIM, J.-O., & MUELLER, C. W. (1978a). Factor analysis: Statistical methods and practical issues. Newbury Park, CA: Sage.
KIM, J.-O., & MUELLER, C. W. (1978b). Introduction to factor analysis: What it is and how to do it. Newbury Park, CA: Sage.
KINDER, B. N. (1992). The problems of R in clinical settings and in research: Suggestions for the future. Journal of Personality Assessment, 58, 252–259.
KING, L. A., & KING, D. W. (1990). Role conflict and role ambiguity: A critical assessment of construct validity. Psychological Bulletin, 107, 48–64.
KING, W. L., & SEEGMILLER, B. (1973). Performance of 14- to 22-month-old black, first-born male infants on two tests of cognitive development: The Bayley Scales and the Infant Psychological Development Scale. Developmental Psychology, 8, 317–326.
KINSLINGER, H. J. (1966). Application of projective techniques in personnel psychology since 1940. Psychological Bulletin, 66, 134–149.
KIRCHER, J. C., & RASKIN, D. C. (1992). Polygraph techniques: History, controversies, and prospects. In P. Suedfeld & P. E. Tetlock (Eds.), Psychology and social policy (pp. 295–308). New York: Hemisphere.
KIRCHNER, W. K. (1966). A note on the effect of privacy in taking typing tests. Journal of Applied Psychology, 50, 373–374.
KIRNAN, J. P., & GEISINGER, K. F. (1986). Review of the General Aptitude Test Battery. In D. J. Keyser & R. C. Sweetland (Eds.), Test critiques (Vol. 5, pp. 150–167). Kansas City, MO: Test Corporation of America.
KIRSCH, I. S., JUNGEBLUT, A., JENKINS, L., & KOLSTAD, A. (1993). Adult literacy in America: A first look at the results of the National Adult Literacy Survey. Washington, DC: US Department of Education.
KITAYAMA, S., & MARCUS, H. R. (Eds.) (1994). Emotion and culture: Empirical studies of mutual influences. Washington, DC: American Psychological Association.

KLEIGER, J. H. (1992). A conceptual critique of the EA:es comparison in the Comprehensive Rorschach System. Psychological Assessment, 4, 288–296.
KLEINMUNTZ, B. (1990). Why we still use our heads instead of formulas: Toward an integrative approach. Psychological Bulletin, 107, 296–310.
KLINE, P. (1993). An easy guide to factor analysis. New York: Routledge.
KLINE, R. B. (1994). Test review: New objective rating scales for child assessment. I. Parent-and teacher-informant inventories of the Behavioral Assessment System for Children, the Child Behavior Checklist, and the Teacher Report Form. Journal of Psychoeducational Assessment, 12, 289–306.
KLINE, R. B., LACHAR, D., & BOERSMA, D. C. (1993). Identification of special education needs with the Personality Inventory for Children (PIC): A hierarchical classification model. Psychological Assessment, 5, 307–316.
KLINE, R. B., LACHAR, D., & GDOWSKI, C. L. (1992). Clinical validity of a Personality Inventory for Children (PIC) profile typology. Journal of Personality Assessment, 58, 591–605.
KLINE, R. B., SNYDER, J., & CASTELLANOS, M. (1996). Lessons from the Kaufman Assessment Battery for Children (K-ABC): Toward a new cognitive assessment model. Psychological Assessment, 8, 7–17.
KLINEBERG, O. (1928). An experimental study of speed and other factors in «racial» differences. Archives of Psychology, No. 93.
KLINGER, E. (1966). Fantasy need achievement as a motivational construct. Psychological Bulletin, 66, 291–308.
KLOPFER, W. G. (1983). Writing psychological reports. In C. E. Walker (Ed.), The handbook of clinical psychology (Vol. 1, pp. 501–527). Homewood, IL: Dow Jones-Irwin.
KLOPFER, W. G., & TAULBEE, E. S. (1976). Projective tests. Annual Review of Psychology, 27, 543–568.
KNAPP, D. J., & CAMPBELL, J. P. (1993). Building a joint-service classification research roadmap: Criterion-related issues (AL/HR-TP-1993-0028). Brooks AFB, TX: Arm-strong Laboratory.
KNAPP, D. J., RUSSELL, T. L., & CAMPBELL, J. P. (1993). Building a joint-service classification research roadmap: Job analysis methodologies (Interim report HumRRO IR-PRD–93–15), Brooks AFB, TX: Armstrong Laboratory.
KNAPP, R. R. (1960). The effects of time limits on the intelligence test performance of Mexican and American subjects. Journal of Educational Psychology, 51, 14–20.
KNOBLOCK, H., & PASAMANICK, B. (1963). Predicting intellectual potential in infancy. American Journal of Diseases of Children, 106, 43–51.
KNOBLOCH, H., & PASAMANICK, B. (1966). Prospective studies on the epidemiology of reproductive casualty: Methods, findings, and some implications. Merrill-Palmer Quarterly, 12, 27–43.
KNOBLOCH, H., & PASAMANICK, B. (Eds.). (1974). Gessell and Amatruda's developmental diagnosis (3rd ed.). New York: Harper & Row.
KNOBLOCH, H., STEVENS, F., & MALONE, A. F. (1980). Manual of developmental diagnosis: The administration and interpretation of revised Gesell and Amatruda Developmental and Neurologic Examination. Philadelphia: Harper & Row.
KNOELL, M., & HARRIS, C. W. (1952). A factor analysis of spelling ability. Journal of Educational Research, 46, 95–111.
KNOFF, H. M. (1989). Review of the Personality Inventory for Children, Revised Format. Tenth Mental Measurements Yearbook, 625–630.
KNOFF, H. M. (1990). Evaluation of projective drawings. In C. R. Reynolds & R. W. Kamphaus (Eds.), Handbook of psychological assessment of children: Personality, behavior, and context (pp. 89–146). New York: Guilford Press.
KNOFF, H. M. (1992). Assessment of social-emotional functioning and adaptive behavior. In E. Vazquez Nutall, I. Romero, & J. Kalesnik (Eds.), Assessing and screening preschoolers: Psychological and educational dimensions (pp. 121–143). Boston: Allyn & Bacon.
KNOFF, H. M. (1993). The utility of human figure drawings in personality and intellectual assessment: Why ask why? School Psychology Quarterly, 8, 191–196.
KNOX, H. A. (1914). A scale based on the work at Ellis Island for estimating mental defect. Journal of the American Medical Association, 62, 741–747.
KOCH, H. L. (1966). Twins and twin relations. Chicago: University of Chicago Press.
KOGAN, N. (1976). Cognitive styles in infancy and early childhood. Hillsdale, NJ: Erlbaum.
KOGAN, N., & BLOCK, J. (1991). Field dependence-independence from early childhood through adolescence: Personality and socialization aspects. In S. Wagner & J. Demick (Eds.), Field dependence-independence: Cognitive style across the life span (pp. 177–207). Hillsdale, NJ: Erlbaum.
KOLB, B., & WHISHAW, I. Q. (1990). Fundamentals of human neuropsychology (3rd ed.). New York: Freeman.
KOPPITZ, E. M. (1964). The Bender Gestalt Test for young children. Orlando, FL: Grune & Stratton.
KOPPITZ, E. M. (1968). Psychological evaluation of children's human figure drawings. Boston: Allyn & Bacon.
KOPPITZ, E. M. (1975). The Bender Gestalt Test for young children: Research and application, 1963–1973. Orlando, FL: Grune & Stratton.
KOPPITZ, E. M. (1984). Psychological evaluation of human figure drawings by middle school pupils. Orlando, FL: Grune & Stratton.
KOTSONIS, M. E., & PATTERSON, C. J. (1980). Comprehension-monitoring skills in learning-disabled children. Developmental Psychology, 16, 541–542.
KOZLOWSKI, S. W., J., KIRSCH, M. P., & CHAO, G. T. (1986). Job knowledge, ratee familiarity, and halo effect: An exploration. Journal of Applied Psychology, 71, 45–49.
KRAEPELIN, E. (1892). Uber die Beeinflassung einfacher psychischer Vorgange durch einige Arzneimittel. Jena: Fischer.
KRAEPELIN, E. (1895). Der psychologische Versuch in der Psychiatrie. Psychologische Arbeiten, 1, 1–91.
KRALL, V. (1986). Projective play techniques. In A. I. Rabin (Ed.), Projective techniques for adolescents and children (pp. 264–278). New York: Springer.
KRAMER, J. H. (1990). Guidelines for interpreting WAIS-R subtest scores. Psychological Assessment, 2, 202–205.
KRAMER, J. H. (1993). Interpretation of individual subtest scores on the WISC-III. Psychological Assessment, 5, 193–196.
KRAMER, J. J., & MITCHELL, J. V., JR. (Eds.). (1985). Computer-based assessment and interpretation: Prospects, promise, and pitfalls [Special issue]. Computers in Human Behavior, 1 (3/4).

KRATOCHWILL, T. R., DOLL, E. J., & DICKSON, W. P. (1991). Use of computer technology in behavioral assessments. In T. B. Gutkin & S. L. Wise (Eds.), The computer and the decision-making process (pp. 125–154). Hillsdale, NJ: Erlbaum.
KRAVETS, M., & WAX, I. (1992). The K & W guide: Colleges and the learning disabled student. New York: Harper Collins.
KROGER, R. O., & WOOD, L. A. (1993). Reification, «faking,» and the Big Five. American Psychologist, 48, 1297–1298.
KRUG, S. E. (Ed.). (1988). Psychware sourcebook (3rd ed.). Kansas City, MO: Test Corporation of America.
KRUG, S. E. (Ed.). (1993). Psychware sourcebook (4th ed.). Champaign, IL: Metritech.
KRUGLANSKI, A. W. (1989). The psychology of being right: The problem of accuracy in social perception and cognition. Psychological Bulletin, 106, 395–409.
KRUMBOLTZ, J. D. (1991). Manual for the Career Beliefs Inventory. Palo Alto, CA: Consulting Psychologists Press.
KUDER, F. E (1966). The Occupational Interest Survey. Personnel and Guidance Journal, 45, 72–77.
KUDER, F. E, & DIAMOND, E. E. (1979). Kuder Occupational Interest Survey: General manual. Chicago: Science Research Associates.
KUDER, F. E, & RICHARDSON, M. W. (1937). The theory of estimation of test reliability. Psychometrika, 2, 151–160.
KUDER, E, & ZYTOWSKI, D. G. (1991). Kuder Occupational Interest Survey Form DD: General manual (3rd ed.), Monterey, CA: CTB Macmillan/McGraw-Hill.
KUHLMANN, F. (1912). A revision of the Binet-Simon system for measuring the intelligence of children. Journal of Psycho-Asthenics, Monograph Supplement, 1, 1–41.
KULIKOWICH, J. M., & ALEXANDER, A. (1994). Evaluating students' errors on cognitive tasks: Applications of polytomous item response theory and log-linear modeling. In C. R. Reynolds (Ed.), Cognitive assessment: A multidisciplinary perspective (pp. 137–154). New York: Plenum Press.
KUMMEROW, J. M. (Ed.).(1991). New directions in career planning and the workplace: Practical strategies for counselors. Palo Alto, CA: Davies-Black.
KURTZ, A. K. (1948). A research test of the Rorschach test. Personnel Psychology, 1, 41–51.
LACHAR, D. (1982). Personality Inventory for Children (PIC): Revised format manual supplement, Los Angeles: Western Psychological Services.
LACHAR, D., & GDOWSKI, C. L. (1979). Actuarial assessment of child and adolescent personality: An interpretive guide for the Personality Inventory for Children profile. Los Angeles: Western Psychological Services.
LACHAR, D., & GRUBER, C. P. (1993). Development of the Personality Inventory for Youth: A self-report companion to the Personality Inventory for Children. Journal of Personality Assessment, 61, 81–98.
LACHAR, D., & GRUBER, C. P. (1995a). Personality Inventory for Youth (PIY) manual: Administration and scoring guide. Los Angeles: Western Psychological Services.
LACHAR, D., & GRUBER, C. P. (1995b). Personality Inventory for Youth (PIY) manual: Technical guide. Los Angeles: Western Psychological Services.
LaDUCA, A. (1994). Validation of professional licensure examinations: Professions theory, test design, and construct validity. Evaluation & the Health Professions, 17, 178–197.
LaFAVE, L. (1966). Essay vs. multiple-choice: Which test is preferable? Psychology in the Schools, 3, 65–69.
LAH, M. I. (1989). Sentence completion tests. In C. S. Newmark (Ed.), Major psychologycal assessment instruments (Vol. 2, pp. 133–163). Boston: Allyn & Bacon.
LALLI, J. S., & GOH, H. (1993). Naturalistic observations in community settings. In J. Reichle & D. P. Wacker (Eds.), Communicative alternatives to challenging behavior: integrating functional assessment and intervention strategies (pp. 11–39). Baltimore: Paul H. Brookes.
LAMBERT, N. M. (1990). Consideration of the Das-Naglieri Cognitive Assessment System. Journal of Psychoeducational Assessment, 8, 338–343.
LAMBERT, N. (1991). The crisis in measurement literacy in psychology and education. Educational Psychologist, 26, 23–35.
LAMBERT, N., NIHIRA, K., & LELAND, H. (1993). AAMR Adaptive Behavior Scale-School-Second Edition: Examiner's manual. Austin, TX: PRO-ED.
LANDFIELD, A. W., & EPTING, R. R. (1987). Personal construct psychology: Clinical and personality assessment. New York: Human Sciences Press.
LANDY, R. J., & FARR, J. L. (1980). Performance rating. Psychological Bulletin, 87, 72–107.
LANDY, F. J., & FARR, J. L. (1983). The measurement of work performance. New York: Academic Press.
LANDY, F. J., SHANKSTER, L.J., & KOHLER, S. S. (1994). Personnel selection and placement. Annual Review of Psychology, 45, 261–296.
LANDY, F., SHANKSTER-CAWLEY, L., & KOHLER MORAN, S. (1995). Advancing personnel selection and placement methods. In A. Howard (Ed.), The changing nature of work (pp. 252–289). San Francisco: Jossey-Bass.
LANG, W. S. (1992). Review of the TEMAS (Tell-Me-A-Story). Eleventh Mental Measurements Yearbook, 925–926.
LANNING, K. (1991). Consistency, scalability, and personality measurement. New York: Springer-Verlag.
LANYON, R. I. (1966). A free-choice version of the EPPS. Journal of Clinical Psychology, 22, 202–205.
LANYON, R. I. & GOODSTEIN, L. D. (1997). Personality assessment (3rd ed.). New York: Wiley.
LAOSA, L. M., SWARTZ, J. D., & DIAZ-GUERRERO, R. (1974). Perceptual-cognitive and personality development of Mexican and Anglo-American children as measured by human figure drawings. Developmental Psychology, 10, 131–139.
LARKIN, J. H., McDERMOTT, J., SIMON, D. F., & SIMON, H. A. (1980a). Expert and novice performance in solving physics problems. Science, 208, 1335–1342.
LARKIN, J. H., McDERMOTT, J., SIMON, D. F., & SIMON, H. A. (1980b). Models of competence in solving physics problems. Cognitive Science, 4, 317–345.
LARKIN, K. C., & WEISS, D. J. (1974). An empirical investigation of computer-administered pyramidal ability testing (Res. Rep. 74–3). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.
LAST, J., & BRUHN, A. R. (1991). The Comprehensive Early Memories Scoring System — Revised. 17 pages. Available from the second author.

LAURENT, J., SWERDLIK, M., & RYBURN, M. (1992). Review of validity research on the Stanford-Binet Intelligence Scale: Fourth Edition. Psychological Assessment, 4, 102–112.

LAUTENSCHLAGER, G. J. (1994). Accuracy and faking of background data. In G. S. Stokes, M. D. Mumford, & W. A. Owens (Eds.), Biodata handbook: Theory, research, and use of biographical information in selection and performance prediction (pp. 391–419). Palo Alto, CA: Consulting Psychologists Press.

LAVE, J. (1988). Cognition in practice: Mind, mathematics, and culture in everyday life. Cambridge, England: Cambridge University Press.

LAWRENCE, S. W., JR. (1962). The effects of anxiety, achievement motivation, and task importance upon performance on an intelligence test. Journal of Educational Psychology, 53, 150–156.

LAZARUS, A. A. (1981). The practice of multimodal therapy. New York: McGraw-Hill.

LEARK, R. A., DUPUY, T. R., GREENBERG, L. M., CORMAN, C. L., & KINDSCHI, C. (1996). T.O.V.A. Test of Variables of Attention : Professional manuals Version 7.0. Los Alamitos, CA: University Attention Disorders.

LECKLITER, I. N., MATARAZZO, J. D., & SILVERSTEIN, A. B. (1986). A literature review of factor analytic studies of the WAIS'R. Journal of Clinical Psychology, 42, 332–342.

LEE, R., & FOLEY, P. P. (1986). Is the validity of a test constant throughout the score range? Journal of Applied Psychology, 71, 641–644.

LEE, Y., JUSSIM, L. J., & McCAULEY, C. R. (1995). Stereotype accuracy: Toward appreciating group differences. Washington, DC: American Psychological Association.

LEFCOURT, H. M. (1991). Locus of control. In J. P. Robinson, PR. Shaver, & L. S. Wrightsman (Eds.), Measures of personality and social psychological attitudes (pp. 413–499). San Diego, CA: Academic Press.

LEFCOURT, H. M., VON BAEYER, C. L., WARE, E. E., & COX, D. V. (1979). The multidimensional-multiattributional causality scale: The development of a goal specific locus of control scale. Canadian Journal of Behavioural Science, 11, 286–304.

LEICHSENRING, R (1991). Discriminating schizophrenics from borderline patients: Study with the Holtzman Inkblot Technique. Psychopathology, 24, 225–231.

LENNEY, E. (1991). Sex roles: The measurement of masculinity, femininity, and androgyny. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), Measures of personality and social psychological attitudes (pp. 573–660). San Diego, CA: Academic Press.

LENNON, R. T. (1966a). A comparison of results of three intelligence tests. In C. I. Chase & H. G. Ludlow (Eds.), Readings in educational and psychological measurement (pp. 198–205). Boston: Houghton Mifflin.

LENNON, R. T. (1966b). Norms: 1963. In A. Anastasi (Ed.), Testing problems in perspective (pp. 243–250). Washington, DC: American Council on Education.

LENS, W., ATKINSON, J. W., & YIP, A. G. (1979). Academic achievement in high school related to «intelligence» and motivation as measured in sixth, ninth, and twelfth grade boys and girls. Unpublished manuscript, University of Michigan, Ann Arbor.

LERNER, B. (1980a). Employment discrimination: Adverse impact, validity, and equality. In P. B. Kurland & G. Casper (Eds.), 1979 Supreme Court Review (pp. 17–49). Chicago: University of Chicago Press.

LERNER, B. (1980b). The war on testing: Detroit Edison in perspective. Personnel Psychology, 33, 11–16.

LERNER, P. M. (1991). Psychoanalytic theory and the Rorschach. New York: Analytic Press.

LERNER, P. M. (1994). Current status of the Rorschach. Contemporary Psychology, 39, 724–725.

LEVIN, J. D. (1992). Theories of the self. Washington, DC: Hemisphere.

LEVY, L. (1963). Psychological interpretation. New York: Holt, Rinehart & Winston.

LEWIS, M. (1973). Infant intelligence tests: Their use and misuse. Human Development, 16, 108–118.

LEWIS, M. (1976). What do we mean when we say «infant intelligence scores?» A sociopolitical question. In M. Lewis (Ed.), Origins of intelligence: Infancy and early childhood (pp. 1–17). New York: Plenum Press.

LEWIS, M., & McGURK, H. (1972). Evaluation of infant intelligence: Infant intelligence scores—true or false? Science, 178 (4066), 1174–1177.

LEZAK, M. D. (1995). Neuropsychological assessment (3rd ed.). New York: Oxford University Press.

LIBEN, L. S. (Ed.). (1983). Piaget and the foundations of knowledge (The Jean Piaget Symposium Series, No. 10). Hillsdale, NJ: Erlbaum.

LIDZ, C. S. (1981). Improving assessment of schoolchildren. San Francisco: Jossey-Bass.

LIDZ, C. S. (Ed.). (1987). Dynamic assessment: An interactive approach to evaluating learning potential. New York: Guilford Press.

LIDZ, C. S. (1991). Practitioners guide to dynamic assessment. New York: Guilford Press.

LIDZ, C. S. (1995). Dynamic assessment and the legacy of L. S. Vygotsky. School Psychology International, 16, 143–153.

LIDZ, C. S. (1997). Dynamic assessment approaches. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), Contemporary intellectual assessment: Theories, tests, and issues (pp. 281–296). New York: Guilford.

LIKERT, R. (1932). A technique for the measurement of attitudes. Archives of Psychology, No. 140.

LIKERT, R., & QUASHA, W. H. (1995). Revised Minnesota Paper Form Board Test: Manual (2nd ed.). San Antonio, TX: Psychological Corporation.

LIM, R. G., & DRASGOW, F. (1990). Evaluation of two methods for estimating item response theory parameters when assessing differential item functioning. Journal of Applied Psychology, 75, 164–174.

LINDEN, M. J., & WHIMBEY, A. (1990). Analytical writing and thinking: Facing the tests. Hillsdale, NJ: Erlbaum.

LINDSLEY, D. B. (1955). The psychology of lie detection. In G. J. Dudycha et al., (Eds.) Psychology for law enforcement officers (chap. 4). Springfield, IL: Charles C Thomas.

LINDZEY, G. (1977). Projective techniques and cross-cultural research. New York: Irvington. (Original work published 1961)

LINDZEY, G., & HERMAN, P. S. (1955). Thematic Apperception Test: A note on reliability and situational validity. Journal of Projective Techniques, 19, 36–42.

LINN, R. L. (1975). Test bias and the prediction of grades in law school. Journal of Legal Education, 27, 293–323.

LINN, R. L. (1978). Single-group validity, differential validity, and differential prediction. Journal of Applied Psychology, 63, 507–512.

LINN, R. L. (1989). Review of the Boehm Test of Basic Concepts—Revised. Tenth Mental Measurements Yearbook, 99–101.
LINN, R. L., & DRASGOW, F. (1987). Implications of the Golden Rule settlement for test construction. Educational Measurement: Issues & Practice, 6, 13–17.
LINN, R. L., & GRONLUND, N. E. (1995). Measurement and assessment in teaching (7th ed.). Upper Saddle River, NJ: Prentice Hall.
LINN, R. L., & WERTS, C. E. (1971). Considerations for studies of test bias. Journal of Educational Measurement, 8, 1–4.
LIPGAR, R. M. (1992). The problem of R in the Rorschach: The value of varying responses. Journal of Personality Assessment, 58, 223–230.
LIPSEY, M. W., & WILSON, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. American Psychologist, 48, 1181–1209.
LITTELL, W. M. (1960). The Wechsler Intelligence Scale for Children: Review of a decade of research. Psychological Bulletin, 57, 132–156.
LIVINGSTON, S. A., & ZIEKY, M. J. (1982). Passing scores: A manual for setting standards of performance on educational and occupational tests. Princeton, NJ: Educational Testing Service.
LOBELLO, S. G., & GULGOZ, S. (1991). Factor analysis of the Wechsler Preschool and Primary Scale of Intelligence—Revised. Psychological Assessment, 3, 130–132.
LOEHLIN, J. C. (1992). Latent variable models: An introduction to factor, path, and structural analysis (2nd ed.). Hillsdale, NJ: Erlbaum.
LOEHLIN, J., LINDZEY, G., & SPUHLER, J. N. (1975). Race, differences in intelligence. New York: Freeman.
LOEVINGER, J. (1966a). The meaning and measurement of ego development. American Psychologist, 21, 195–206.
LOEVINGER, J. (1966b). A theory of test response. In A. Anastasi (Ed.), Testing problems in perspective (pp. 545–556). Washington, DC: American Council on Education.
LOEVINGER, J. (1976). Ego development. San Francisco: Jossey-Bass.
LOEVINGER, J. (1985). Revision of the Sentence Completion Test for ego development. Journal of Personality and Social Psychology, 48, 420–427.
LOEVINGER, J. (1987). Paradigms of personality. New York: Freeman.
LOEVINGER, J. (1993). Measurement of personality: True or false? Psychological Inquiry, 4, 1–16.
LOEVINGER, J. (1994). Has psychology lost its conscience? Journal of Personality Assessment, 62, 2–8.
LOEVINGER, J., & OSSORIO, A. G. (1958). Evaluation in therapy by self-report: A paradox. American Psychologist, 13, 366.
LOEVINGER, J., & WESSLER, R. (1970). Measuring ego development: Vol. 1. Construction and use of a sentence completion test. San Francisco: Jossey-Bass.
LOEVINGER, J., WESSLER, R., & REDMORE, C. (1970). Measuring ego development: Vol. 2. Scoring manual for women and girls. San Francisco: Jossey-Bass.
LOFTUS, E. F. (1993). The reality of repressed memories. American Psychologist, 48, 518–537.
LOKAN, J. J., & TAYLOR, K. F. (Eds.). (1986). Holland in Australia: A vocational choice theory in research and practice. Melbourne: Australian Council for Educational Research.
LONNER, W. J., & ADAMS, H. L. (1972). Interest patterns of psychologists in nine Western nations. Journal of Applied Psychology, 56. 141–151.
LONNER, W. J., & BERRY, J. W. (Eds.). (1986). Field methods in cross-cultural research. Beverly Hills, CA: Sage.
LORD, F. M. (1952). The relation of the reliability of multiple-choice tests to the distribution of item difficulties. Psychometrika, 17, 181–194.
LORD, F. M. (1970). Some test theory for tailored testing. In W. H. Holtzman (Ed.), Computer-assisted instruction, testing, and guidance (pp. 139–183). New York: Harper & Row.
LORD, F. M. (1971a). The self-scoring flexilevel test. Journal of Educational Measurement, 8, 147–151.
LORD, F. M. (1971b). A theoretical study of the measurement effectiveness of flexilevel tests. Educational and Psychological Measurement, 31, 805–813.
LORD, F. M. (1971c). A theoretical study of two-stage testing. Psychometrika, 36, 227–241.
LORD, F. M. (1980). Applications of item response theory to practical testing problems. Hills-dale, NJ: Erlbaum.
LORET, P. G., SEDER, A., BIANCHINI, J. C., & VALE, C. A. (1974). Anchor Test Study: Equivalence and norms tables for selected reading achievement tests. Washington, DC: U.S. Government Printing Office.
LORGE, I. (1945). Schooling makes a difference. Teachers College Record, 46, 483–492.
LOWMAN, R. L. (1989). Pre-employment screening for psychopathology: A guide to professional practice. Sarasota, FL: Professional Resource Press.
LOWMAN, R. L. (1991). The clinical practice of career assessment: Interests, abilities, and personality. Washington, DC: American Psychological Association.
LOWMAN, R. L. (1993). Counseling and psychotherapy of work dysfunctions. Washington, DC: American Psychological Association.
LOYD, B. H. (1995). Review of the Family Environment Scale, Second Edition. Twelfth Mental Measurements Yearbook, 385–386.
LU, C., & SUEN, H. K. (1995). Assessment approaches and cognitive styles. Journal of Educational Measurement, 32, 1–17.
LUBIN, B., LARSEN, R. M., & MATARAZZO, J. D. (1984). Patterns of psychological test usage in the United States: 1935–1982. American Psychologist, 39, 451–454.
LUBINSKI, D., & BENBOW, C. P. (1995). An opportunity for empiricism [Review of Multiple intelligences: The theory and practice]. Contemporary Psychology, 40, 935–940.
LUBINSKI, D., & DAWIS, R. V. (1992). Aptitudes, skills, and proficiencies. In M. D. Dunnette & L. M. Hough (Eds.), Handbook of industrial and organizational psychology (2nd ed., Vol. 3, pp. 1–59). Palo Alto, CA: Consulting Psychologists Press.
LUKAS, S. (1993). Where to start and what to ask: An assessment handbook. New York: W. W. Norton.
LUKHELE, R., THISSEN, D., & WAINER, H. (1994). On the relative value of multiple-choice, constructed response, and examinee-selected items on two achievement tests. Journal of Educational Measurement, 31, 234–250.

LURIA, A. R. (1966). Human brain and psychological processes. New York: Harper & Row.
LURIA, A. R. (1973). The working brain. New York: Basic Books.
LURIA, A. R. (1980). Higher cortical functions in man (2nd ed.). New York: Basic Books.
LUTEY, C., & COPELAND, E. P. (1982). Cognitive assessment of the school-age child. In C. R. Reynolds & T. B. Gutkin (Eds.), The handbook of school psychology (pp. 121–155). New York: Wiley.
LYKKEN, D. T. (1981). A tremor in the blood: Uses and abuses of the lie detector test. New York: McGraw-Hill.
LYKKEN, D. T. (1992). Controversy: The fight-or-flight response in Homo scientificus. In P. Suedfeld & P. E. Tetlock (Eds.), Psychology and social policy (pp. 309–325). New York: Hemisphere.
LYON, M. A., & MACDONALD, N. T. (1990). Academic self-concept as a predictor of achievement for a sample of elementary school students. Psychological Reports, 66, 1135–1142.
MABRY, L. (1995). Review of the Metropolitan Readiness Tests, Fifth Edition. Twelfth Mental Measurements Yearbook, 611–612.
MacCALLUM, R. C., & BROWNE, M. W. (1993). The use of causal indicators in covariance structure models: Some practical issues. Psychological Bulletin, 114, 533–541.
MacCALLUM, R. C., WEGENER, D. T, UCHINO, B. N., & FABRIGAR, L. R. (1993). The problem of equivalent models in applications of covariance structure analysis. Psychological Bulletin, 114, 185–189.
MACHOVER, K. (1949). Personality projection in the drawing of the human figure. A method of personality investigation. Springfield, IL: Charles C Thomas.
MacLENNAN, R. N. (1992). Personality Research Form (PRF): Annotated research bibliography with author and subject indexes. Port Huron, MI: Sigma Assessment Systems.
MACMANN, G. M., & BARNETT, D. W. (1994a). Some additional lessons from the Wechsler scales: A rejoinder to Kaufman and Keith. School Psychology Quarterly, 9, 223–236.
MACMANN, G. M., & BARNETT, D. W. (1994b). Structural analysis of correlated factors: Lessons from verbal-performance dichotomy of the Wechsler scales. School Psychology Quarterly, 9, 161–197.
MacMILLAN, D. L., GRESHAM, F. M., & SIPERSTEIN, G. N. (1993). Conceptual and psychometric concerns about the 1992 AAMR definition of mental retardation. American Journal on Mental Retardation, 98, 325–335.
MADDI, S. R. (1989). Personality theories: A comparative analysis (5th ed.). Chicago: Dorsey Press.
MADDUX, J. E. (Ed.). (1995). Self-efficacy, Adaptation, and adjustment: Theory, research, and application. New York: Plenum Press.
MAEL, F. A. (1991). A conceptual rationale for the domain and attributes of biodata items. Personnel Psychology, 44, 763–792.
MAHONEY, M. J. (1991). Human change processes: The scientific foundations of psychotherapy. New York: Basic Books.
MAHURIN, R. K. (1992). Review of the Computer Programmer Aptitude Battery. Eleventh Mental Measurements Yearbook, 225–227.
MAIER, M. H. (1972). Effects of educational level on prediction of training success with ACB (Tech. Res. Note 225). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
MAIER, M. H., & FUCHS, E. F. (1973). Effectiveness of selection and classification testing (Res. Rep. 1179). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
MAIER, M. H., & HIRSHFELD, S. F. (1978). Criterion-referenced job proficiency testing: A large scale application (Res. Rep. 1193). Alexandria, VA: U. S. Army Research Institute for the Behavioral and Social Sciences.
MALGADY, R. G., ROGLER, L. H., & COSTANTINO, G. (1987). Ethnocultural and linguistic bias in mental health evaluation of Hispanics. American Psychologist, 42, 228–234.
MALLER, S. J., & BRADEN, J. P. (1993). The construct and criterion-related validity of the WISC-III with deaf adolescents. Journal of Psychoeducational Assessment, WISC-III Monograph, 105–113.
MALONEY, M. P., & WARD, M. P. (1976). Psychological assessment: A conceptual approach. New York: Oxford University Press.
MANDLER, G., & SARASON, S. B. (1952). A study of anxiety and learning. Journal of Abnormal and Social Psychology, 47, 166–173.
MANOLEAS, P. (Ed.). (1995). The cross-cultural practice of clinical case management. Binghamton, NY: Haworth Press.
MANUELE-ADKINS, C. (1989). Review of The Self-Directed Search: A guide to educational and vocational planning—1985 Revision. Tenth Mental Measurements Yearbook, 738–740.
MARCO, G. L. (1992). Review of the Computer Literacy and Computer Science Tests. Eleventh Mental Measurements Yearbook, 220–222.
MARIN, G., & MARIN, B. V. (1991). Research with Hispanic populations. Newbury Park: CA: Sage.
MARKS, P. A., SEEMAN, W., & HALLER, D. L. (1974). The actuarial use of the MMPI with adolescents and adults. Baltimore: Williams & Wilkins.
MARKUS, H., & WURF, E. (1987). The dynamic self-concept: A social psychological perspective. Annual Review of Psychology, 38, 299–337.
MARSH, D. T, LINBERG, L. M., & SMELTZER, J. K. (1991). Human figure drawings of adjudicated and nonadjudicated adolescents. Journal of Personality Assessment, 57, 77–86.
MARSH, H. W. (1990a). Causal ordering of academic achievement: A multiwave, longitudinal panel analysis. Journal of Educational Psychology, 82, 646–656.
MARSH, H. W., (1990b). The structure of academic self-concept: The Marsh/Shavelson model. Journal of Educational Psychology, 82, 623–636.
MARSH, H. W., BYRNE, B. M., & SHAVELSON, R. J. (1992). A multidimensional, hierarchical self-concept. In T. M. Brinthaupt & R. P. Lipka (Eds.), The self: Definitional and methodological issues (pp. 44–95). Albany: State University of New York Press.
MARSH, H. W., & SHAVELSON, R. (1985). Self-concept: Its multifaceted, hierarchical structure. Educational Psychologist, 20, 107–123.
MARTIN, S. L., AND RAJU, N. S. (1992). Determining cutoff scores that optimize utility: A recognition of recruiting costs. Journal of Applied Psychology, 77, 15–23.

MARUISH, M. E. (Ed.). (1994). The use of psychological testing for treatment planning and outcome assessment. Hillsdale, NJ: Erlbaum.
MARUISH, M. E., & MOSES, J. A. (Eds.). (1997). Clinical neuropsychology: Theoretical foundations for practitioners. Mahwah, NJ: Erlbaum.
MASH, E. J., & TERDAL, L. G. (Eds.). (1988). Behavioral assessment of childhood disorders: Selected core problems (2nd ed.). New York: Guilford Press.
MASLING, J. (1959). The effects of warm and cold interaction on the administration and scoring of an intelligence test. Journal of Consulting Psychology, 23, 336–341.
MASLING, J. (1960). The influences of situational and interpersonal variables in projective testing. Psychological Bulletin, 57, 65–85.
MASLING, J. (1965). Differential indoctrination of examiners and Rorschach responses. Journal of Consulting Psychology, 29, 198–201.
MATARAZZO, J. D. (1972). Wechsler's measurement and appraisal of adult intelligence (5th ed.). Baltimore: Williams & Wilkins.
MATARAZZO, J. D. (1983). Computerized psychological testing. Science, 221, 323.
MATARAZZO, J. D. (1986a). Computerized clinical psychological test interpretation: Unvalidated plus all mean and no sigma. American Psychologist, 41, 14–24.
MATARAZZO, J. D. (1986b). Response to Fowler and Butcher on Matarazzo. American Psychologist, 41, 96.
MATARAZZO, J. D. (1990). Psychological assessment versus psychological testing: Validation from Binet to the school, clinic, and courtroom. American Psychologist, 45, 999–1017.
MATSON, J. L. (1995). Comments on Gresham, MacMillan, and Siperstein's paper 'Critical analysis of the 1992 AAMR definition: Implications for school psychology'. School Psychology Quarterly, 10, 20–23.
MATTHEWS, G., JONES, D. M., & CHAMBERLAIN, A. G. (1992). Predictors of individual differences in mail-coding skills and their variation with ability level. Journal of Applied Psychology, 77, 406–418.
MAY, T. M. (1990). An evolving relationship. Counseling Psychologist, 18, 266–270.
MAYER, J. D., & SALOVEY, P. (1993). The intelligence of emotional intelligence. Intelligence, 17, 433–442.
MAZE, M., & MAYALL, D. (Eds.). (1995). The enhanced guide for occupational exploration. Indianapolis, IN: JIST.
MAZZEO, J., DRUESNE, B., RAFFELD, P. C., CHECKETTS, K. T., & MUHLSTEIN, A. (1991). Compatibility of computer and paper-and-pencil scores for two CLEP general examinations (College Board Rep. No. 91–5; ETS Res. Rep. No. 92–14). Princeton, NJ: Educational Testing Service.
McALLISTER, L. W. (1996). A practical guide to CPI interpretation (3rd ed.). Palo Alto, CA: Consulting Psychologists Press.
McANDREW, F. T. (1993). Environmental psychology. Pacific Grove, CA: Brooks/Cole.
McARTHUR, D. S., & ROBERTS, G. E. (1982). Roberts Apperception Test for Children: Manual. Los Angeles: Western Psychological Services.
McBRIDE, J. R., & MARTIN, J. T. (1983). Reliability and validity of adaptive tests in a military setting. In. D. J. Weiss (Ed.), New horizons in testing (pp. 223–236). Orlando, Fl: Academic Press.
McCALL, R. B. (1976). Toward an epigenetic conception of mental development in the first three years of life. In M. Lewis (Ed.), Origins of intelligence: Infancy and early childhood (pp. 97–121). New York: Plenum Press.
McCALL, R. B. (1981). Nature-nurture and the two realms of development: A proposed integration with respect to mental development. Child Development, 52, 1–12.
McCALL, R. B., APPELBAUM, M. I., & HOGARTY, P. S. (1973). Developmental changes in mental performance. Monographs of the Society for Research in Child Development, 38 (3, Serial No. 150).
McCALL, R. B., EICHORN, D. H., & HOGARTY, P. S. (1977). Transitions in early mental development. Monographs of the Society for Research in Child Development, 42 (3, Serial No. 171).
McCALL, R. B., HOGARTY, P. S., & HURLBURT, N. (1972). Transitions in infant sensorimotor development and the prediction of childhood IQ. American Psychologist, 27, 728–748.
McCALL, W. A. (1922). How to measure in education. New York: Macmillan.
McCALLUM, R. S. (1985). Review ofPeabody Picture Vocabulary Test — Revised. Ninth Mental Measurements Yearbook, Vol. 2, 1126–1127.
McCALLUM, R. S. (1990). Determining the factor structure of the Stanford-Binet: Fourth Edition — The right choice. Journal of Psychoeducational Assessment, 8, 436–442.
McCARDLE, J. J. (1989). A structural modeling experiment with multiple growth functions. In R. Kanfer, P L. Ackerman, & R. Cudek (Eds.), Abilities, motivation, and methodology (pp. 203–237). Hillsdale, NJ: Erlbaum.
McCARTHY, D. (1944). A study of the reliability of the Goodenough drawing test of intelligence. Journal of Psychology, 18, 201–216.
McCARTHY, D. (1972). Manual for the McCarthy Scales of Children's Abilities. New York: Psychological Corporation.
McCLELLAND, D. C. (1966). Longitudinal trends in the relation of thought to action. Journal of Consulting Psychology, 30, 479–483.
McCLELLAND, D. C. (1976). The achieving society. New York: Irvington. (Original work published 1961)
McCLELLAND, D. C. (1985). Human motivation. Glenview, IL: Scott, Foresman.
McCLELLAND, D. C., ATKINSON, J. W., CLARK, R. A., & LOWELL, E. L. (1976). The achievement motive. New York: Irvington. (Original work published 1953)
McCORMICK, E. J. (1979) Job analysis: Methods and applications. New York: AMACOM.
McCORMICK, E. J. (1983). Job and task analysis. In M. D. Dunnette (Ed.), Handbook of industrial and organizational psychology (pp. 651–696). New York: Wiley.
McCORMICK, E. J., & ILGEN, D. (1980). Industrial psychology (7th ed.). Englewood Cliffs, NJ: Prentice Hall.
McCORMICK, E. J., JEANNERET, P. R., & MECHAM, R. C. (1972). A study of job characteristics and job dimensions as based on the Position Analysis Questionnaire (PAQ). Journal of Applied Psychology, 56, 347–368.
McCRAE, R. R., & JOHN, O. P. (1992). An introduction to the five-factor model and its applications. Journal of Personality, 60, 175–215.

MCCUSKER, P. J. (1994). Validation of Kaufman, Ishikuma, and Kaufman-Packer's Wechsler Adult Intelligence Scale — Revised short forms on a clinical sample. Psychological Assessment, 6, 246–248.
MCDANIEL, M. A., WHETZEL, D. L., SCHMIDT, F. L., & MAURER, S. D. (1994). The validity of employment interviews: A comprehensive review and meta-analysis. Journal of Applied Psychology, 79, 599–616.
McDERMOTT, P. A., FANTUZZO, J. W., & GLUTTING, J. J. (1990). Just say no to subtest analysis: A critique on Wechsler theory and practice. Journal of Psychoeducational Assessment, 8, 290–302.
McDERMOTT, P. A., FANTUZZO, J. W., GLUTTING, J. J., WATKINS, M. W., & BAGGALEY, A. R. (1992). Illusions of meaning in the ipsative assessment of children's ability. Journal of Special Education, 25, 504–526.
McDERMOTT, P. A., GLUTTING, J. J., JONES, J. N., & NOONAN, J. V. (1989). Typology and prevailing composition of core profiles in the WAIS-R standardization sample. Psychological Assessment, 1, 118–125.
McDOWELL, C., & ACKLIN, M. W. (1996). Standardizing procedures for calculating Rorschach interrater reliability: Conceptual and empirical foundations. Journal of Personality Assessment, 66, 308–320.
McGEE, M. G. (1979). Human spatial abilities: Psychometric studies and environmental, genetic, hormonal, and neurological influence. Psychological Bulletin, 86, 889–918.
McGREW, K. S. (1994). Clinical interpretation of the Woodcock-Johnson Tests of Cognitive Ability — Revised. Boston: Allyn & Bacon.
McGREW, K. S., WERDER, J. K., & WOODCOCK, R. W. (1991). Woodcock -Johnson: Technical manual. Alien, TX: DLM.
McGREW, M. W., & TEGLASI, H. (1990). Formal characteristics of Thematic Apperception Test stories as indices of emotional disturbance in children. Journal of Personality Assessment, 54, 639–655.
McHENRY, J. J., HOUGH, L. M., TOQUAM, J. L., HANSON, M. A., & ASHWORTH, S. (1990). Project A validity results: The relationship between predictor and criterion domains. Personnel Psychology, 43, 335–354.
McINTYRE, R. M., SMITH, D. E., & HASSETT, C. E. (1984). Accuracy of performance ratings as affected by rater training and perceived purpose of rating. Journal of Applied Psychology, 69, 147–156.
McKENNA, F. P. (1984). Measures of field dependence: Cognitive style or cognitive ability? Journal of Personality and Social Psychology, 47, 593–603.
McKEOWN, B. & THOMAS, D. (1988). Q methodology. Newbury Park, CA: Sage.
McNEELY, S. (1995). Review of the Alcohol Use Inventory. Twelfth Mental Measurements Yearbook, 66–67.
McREYNOLDS, P. (1975). Historical antecedents of personality assessment. In R McReynolds (Ed.), Advances in psychological assessment (Vol. 3, pp. 477–532). San Francisco: Jossey-Bass.
McREYNOLDS, P. (1986). History of assessment in clinical and educational settings. In R. 0. Nelson & S. C. Hayes (Eds.), Conceptual foundations of behavioral assessment (pp. 42–80). New York: Guilford Press.
McREYNOLDS, P., & DEVOGE, S. (1978). Use of improvisational techniques in assessment. In P. McReynolds (Ed.), Advances in psychological assessment (Vol. 4, pp. 222–227). San Francisco: Jossey-Bass.
MEAD, A. D., & DRASGOW, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. Psychological Bulletin, 114, 449–458.
MEEHL, P. E. (1945). An investigation of a general normality or control factor in personality testing. Psychological Monographs, 59 (4, Whole No. 274).
MEEHL, P. E. (1954). Clinical versus statistical prediction: A theoretical analysis and a review of the evidence. Minneapolis: University of Minnesota Press.
MEEHL, P. E. (1956). Wanted—a good cookbook. American Psychologist, 11, 263–272.
MEEHL, P. E. (1992). Factors and taxa, traits and types, differences of degree and differences in kind. Journal of Personality, 60, 117–174.
MEEHL, P. E. (1995). Extension of the MAXCOV-HITMAX taxonomic procedure to situations of sizable nuisance covariance. In D. Lubinski & R. V. Dawis (Eds.), Assessing individual differences in human behavior: New concepts, methods, and findings (pp. 81–92). Palo Alto, CA: Davies-Black.
MEEHL, P. E., & GOLDEN, R. (1982). Taxometric methods. In P. Kendall & J. N. Butcher (Eds.), Handbook of research methods in clinical psychology (pp. 127–181). New York: Wiley.
MEEHL, P. E., & ROSEN, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. Psychological Bulletin, 52, 194–216.
MEEHL, P. E., & YONCE, L. J. (1994). Taxometric analysis: I. Detecting taxonicity with two quantitative indicators using means above and below a sliding cut (MAMBAC procedure). Psychological Reports, 74, 1059–1274.
MEEKER, M., MEEKER, R., & ROID, G. H. (1985). Structure of Intellect Incoming Abilities Test (SOI-LA): Manual. Los Angeles: Western Psychological Services.
MEGARGEE, E. I. (1966). The relation of response length to the Holtzman Inkblot Technique. Journal of Consulting Psychology, 30, 415–419.
MEHRYAR, A. H., TASHAKKORI, A., YOUSEFI, F., & KHAJAVI, F. (1987). The application of the Goodenough-Harris Draw-A-Man Test to a group of Iranian children in the city of Shiraz. British Journal of Educational Psychology, 57, 401–406.
MEIER, M. J. (1985). Review of Halstead-Reitan Neuropsychological Test Battery. Ninth Mental Measurements Yearbook, Vol. 1, 646–649.
MEIER, S. T. (1993). Revitalizing the measurement curriculum: Four approaches for emphasis in graduate education. American Psychologist, 48, 886–891.
MELLENBERGH, G. J. (1994). Generalized linear item response theory. Psychological Bulletin, ii 5, 300–307.
MELOY, J. R., & SINGER, J. (1991). A psychoanalytic view of the Rorschach Comprehensive System «special scores.» Journal of Personality Assessment, 56, 202–217.
MELTZOFF, J. (1951). The effect of mental set and item structure upon responses to a projective test. Journal of Abnormal and Social Psychology, 46, 177–189.
MENNE, J. W., MCCARTHY, W., & MENNE, J. (1976). A systems approach to the content validation of employee selection procedures. Public Personnel Management, 5, 387–396.

MERENDA, R. E. (1995). Substantive issues in the Soroka v. Dayton-Hudson case. Psychological Reports, 77, 595–606.
MERLUZZI, T. V. (1991). Representation of information about self and other: A multidimensional scaling analysis. In M. J. Horowitz (Ed.), Person schemas and maladaptive interpersonal patterns (pp. 155–166). Chicago: University of Chicago Press.
MESSER, D. J., McCARTHY, M. E., McQUISTON, S., MacTURK, R. H., YARROW, L. J., & VIETZE, P. M. (1986). Relation between mastery behavior in infancy and competence in early childhood. Developmental Psychology, 22, 366–372.
MESSER, S. B. (1976). Reflection-impulsivity: A review. Psychological Bulletin, 83, 1026–1052.
MESSICK, S. (1980a). The effectiveness of coaching for the SAT: Review and reanalysis of research from the fifties to the FTC. Princeton, NJ: Educational Testing Service.
MESSICK, S. (1980b). Test validity and the ethics of assessment. American Psychologist, 35, 1012–1027.
MESSICK, S. (1981). The controversy over coaching: Issues of effectiveness and equity. In B. F. Green (Ed.), Issues in testing: Coaching, disclosure, and ethnic bias (pp. 21–53). San Francisco: Jossey-Bass.
MESSICK, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. Braun (Eds.), Test validity (pp. 33–45). Hillsdale, NJ: Erlbaum.
MESSICK, S. (1989). Validity. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 13–103). New York: American Council on Education/Macmillan.
MESSICK, S. (1992). Multiple intelligences or multilevel intelligence? Selective emphasis on distinctive properties of hierarchy: On Gardner's Frames of mind and Sternberg's Beyond IQ in the context of theory and research on the structure of human abilities. Psychological Inquiry, 3 (4), 365–384.
MESSICK, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. American Psychologist, 50, 741–749.
MESSICK, S., et al. (1976). Individuality in learning. San Francisco: Jossey-Bass.
MESSICK, S., BEATON, A., & LORD, F. (1983). National Assessment of Educational Progress reconsidered: A new design for a new era. Princeton, NJ: National Assessment of Educational Progress.
MESSICK, S., & JUNGEBLUT, A. (1981). Time and method in coaching for the SAT. Psychological Bulletin, 89, 191–216.
MEYER, G. J. (1992). Response frequency problems in the Rorschach: Clinical and re-search implications with suggestions for the future. Journal of Personality Assessment, 58, 231–244.
MEYER, G. J. (1993). The impact of response frequency on the Rorschach constellation indices and on their validity with diagnostic and MMPI-2 criteria. Journal of Personality Assessment, 60, 153–180.
MEYER, P., & DAVIS, S. (1992). The CPI applications guide. Palo Alto, CA: Consulting Psychologists Press.
MEYERS, J. F. (1992). Soroka v. Dayton Hudson Corp. — Is the door closing on pre-employment testing of applicants? Employee Relations Law Journal, 17, 645–653.
MIDDLETON, H. A., KEENE, R. G., & BROWN, G. W. (1990). Convergent and discriminant validities of the Scales of Independent Behavior and the Revised Vineland Adaptive Behavior Scales. American Journal on Mental Retardation, 94, 669–673.
MILLER, A. (1991a). Personality types: A modem synthesis. Calgary, Alberta, Canada: University of Calgary Press.
MILLER, A. (1991b). Personality types, learning styles, and educational goals. Educational Psychology, 11, 217–238.
MILLER, L. T., & LEE, C. J. (1993). Construct validation of the Peabody Picture Vocabulary Test-Revised: A structural equation model of the acquisition order of words. Psychological Assessment, 5, 438–441.
MILLER, P. C., LEFCOURT, H. M., & WARE, E. E. (1983). The construction and development of the Miller Marital Locus of Control Scale. Canadian Journal of Behavioural Science, 15, 266–279.
MILLER, R. J. (1973). Cross-cultural research in the perception of pictorial materials. Psychological Bulletin, 80, 135–150.
MILLER, T. L. (Ed.). (1984). Special issue: Kaufman Assessment Battery for Children. Journal of Special Education, 18 (3), 211–444.
MILLER-JONES, D. (1989). Culture and testing. American Psychologist, 44, 360–366.
MILLMAN, J., BISHOP, C. H., & EBEL, R. (1965). An analysis of test-wiseness. Educational and Psychological Measurement, 25, 707–726.
MILLMAN, J., & GREENE, J. (1989). The specification and development of tests of achievement and ability. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 335–366). New York: American Council on Education/Macmillan.
MILLON, T. (1969). Modem psychological pathology: A biosocial approach to maladaptive learning and functioning. Philadelphia: Saunders.
MILLON, T. (1981). Disorders of personality, DSM-III: Axis II. New York: Wiley.
MILLON, T. (1990). Toward a new personology: An evolutionary model. New York: Wiley.
MILLON, T. (1994). Millon Index of Personality Styles (MIPS) manual. San Antonio, TX: Psychological Corporation.
MILLON, T. (with Davis, R. D., and Millon, C. M., Wenger, A., Van Zuilen, M. H., Fuchs, M., & Millon, R. B.). (1996). Disorders of personality: DSM-IV and beyond (2nd ed.). New York: Wiley.
MILLON, T., GREEN, C. J., & MEAGHER, R. B., JR. (1982). Millon Adolescent Personality Inventory manual. Minneapolis, MN: National Computer Systems.
MILLON, T., MILLON, C., & DAVIS, R. (1993). Millon Adolescent Clinical Inventory (MACI) manual. Minneapolis, MN: National Computer Systems.
MILLON, T., MILLON, C., & DAVIS, R. (1994). MCMI-III manual: Millon Clinical Multiaxial Inventory-III. Minneapolis, MN: National Computer Systems.
MISCHEL, W. (1968). Personality and assessment. New York: Wiley.
MISCHEL, W. (1969). Continuity and change in personality. American Psychologist, 24, 1012–1018.
MISCHEL, W. (1973). Toward a cognitive social learning reconceptualization of personality. Psychological Review, 80, 252–283.
MISCHEL, W. (1977). On the future of personality measurement. American Psychologist, 32, 246–254.
MISCHEL, W. (1979). On the interface of cognition and personality: Beyond the person-situation debate. American Psychologist, 34, 740–754.
MISCHEL, W., & PEAKE, P. K. (1982). Beyond deja vu in the search for cross-situational consistency. Psychological Review, 89, 730–755.

MISLEVY, R. J. (1993). A framework for studying differences between multiple-choice and free-response test items. In R. E. Bennett & W. C. Ward (Eds.), Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment (pp. 75–106). Hillsdale, NJ: Erlbaum.
MISTRY, J., & ROGOFF, B. (1985). A cultural perspective on the development of talent. In R. D. Horowitz & M. O'Brien (Eds.), The gifted and talented: Developmental perspectives. Washington, DC: American Psychological Association.
MITCHELL, B. C. (1967). Predictive validity of the Metropolitan Readiness Tests and the Murphy-Durrell Reading Readiness Analysis for white and negro pupils. Educational and Psychological Measurement, 27, 1047–1054.
MITCHELL, T. W., & KLIMOSKI, R. J. (1986). Estimating the validity of cross-validity estimation. Journal of Applied Psychology, 71, 311–317.
MOEN, P., ELDER, G. H., JR., & LUSCHER, K. (Eds.). (1995). Examining lives in context: Perspectives on the ecology of human development. Washington, DC: American Psychological Association.
MOLLENKOPF, W. G. (1950a). An experimental study of the effects on item-analysis data of changing item placement and test time limit. Psychometrika, 15, 291–317.
MOLLENKOPF, W. G. (1950b). Predicted differences and differences between predictions. Psychometrika, 15, 409–417.
MOORE, B. S., & ISEN, A. M. (Eds.).(1990). Affect and social behavior. New York: Cambridge University Press.
MOORE, H. W., & UNSINGER, P. C. (Eds.). (1987). The police assessment center. Springfield, IL: Charles C Thomas.
MOORE, M. S., & MCLAUGHLIN, L. (1992). Assessment of the preschool child with visual impairment. In E. Vazquez Nutall, I. Romero, & J. Kalesnik (Eds.), Assessing and screening preschoolers: Psychological and educational dimensions (pp. 345–368). Boston: Allyn & Bacon.
MOOS, R. H. (1974). Evaluating treatment environments; A social ecological approach. New York: Wiley.
MOOS, R. (1993a). The Family Environment Scale: An annotated bibliography. Palo Alto, CA: Stanford University and VA Medical Center, Center for Health Care Evaluation.
MOOS, R. (1993b). The Group Environment Scale: An annotated bibliography. Palo Alto, CA: Stanford University and VA Medical Center, Center for Health Care Evaluation.
MOOS, R. (1993c). The Work Environment Scale: An annotated bibliography. Palo Alto, CA: Stanford University and VA Medical Center, Center for Health Care Evaluation.
MOOS, R. H. (1994a). The Social Climate Scales: A users guide. Palo Alto, CA: Consulting Psychologists Press.
MOOS, R. H. (1994b). Work Environment Scale manual: Development, applications, research (3rd ed.). Palo Alto, CA: Consulting Psychologists Press.
MOOS, R. H., & MOOS, B. S. (1994). Family Environment Scale manual: Development, applications, research (3rd ed.). Palo Alto, CA: Consulting Psychologists Press.
MOOS, R. H., & SPINRAD, S. (1984). The social climate scales: An annotated bibliography, 1979–1983. Palo Alto, CA: Consulting Psychologists Press.
MORELAND, K. L. (1985). Validation of computer-based test interpretations: Problems and prospects. Journal of Consulting and Clinical Psychology, 53, 816–825.
MORELAND, K. L. (1987). Computer-based test interpretation: Advice to the consumer. Applied Psychology: An International Review, 36 (3/4), 385–399.
MORELAND, K. L. (1992). Computer-assisted psychological assessment. In M. Zeidner & R. Most (Eds.), Psychological testing: An inside view (pp. 343–376). Palo Alto, CA: Consulting Psychologists Press.
MORELAND, K. L., EYDE, L. D., ROBERTSON, G. J., PRIMOFF, E. S., & MOST, R. B. (1995). Assessment of test user qualifications: A research-based measurement procedure. American Psychologist, 50, 14–23.
MORENO, J. L. (1953). Who shall survive? Foundations of sociometry, group psychotherapy, and sociodrama (2nd ed.). New York: Beacon House.
MORENO, K. E., WETZEL, C. D., McBRIDE, J. R., & WEISS, D. J. (1984). Relationship between corresponding Armed Services Vocational Aptitude Battery (ASVAB) and computerized adaptive testing (CAT) subtests. Applied Psychological Measurement, 8, 155–163.
MOREY, L. C. (1991). Personality Assessment Inventory: Professional manual. Odessa, FL: Psychological Assessment Resources.
MORGAN, G. A., & HARMON, R. J. (1984). Developmental transformations in mastery motivation. In R. N. Emde & R. J. Harmon (Eds.), Continuities and discontinuities in development (pp. 263–291). New York: Plenum Press.
MORGAN, W. G. (1995). Origin and history of the Thematic Apperception Test images. Journal of Personality Assessment, 65, 237–254.
MORRIS, J. H., SHERMAN, J. D., & MANSFIELD, E. R. (1986). Failures to detect moderating effects with ordinary least squares-moderated multiple regressions: Some reasons and a remedy. Psychological Bulletin, 99, 282–288.
MORRISON, J. (1995). The first interview: Revised for DSM-IV. New York: Guilford Press.
MORRISON, T. L., EDWARDS, D. W., & WEISSMAN, H. N. (1994). The MMPI and MMPI–2 as predictors of psychiatric diagnosis in an outpatient sample. Journal of Personality Assessment, 62, 17–30.
MOSES, J. L. (1985). Using clinical methods in a high-level management assessment center. In H. J. Bernardin & D. A. Bownas (Eds.), Personality assessment in organizations (pp. 177–192). New York: Praeger.
MOSSHOLDER, K. W., & ARVEY, R. D. (1984). Synthetic validity: A conceptual and comparative review. Journal of Applied Psychology, 69, 322–333.
MUELLER, D. J. (1986). Measuring social attitudes: A handbook for researchers and practitioners. New York: Teachers College Press.
MUELLER, R. O. (1995). Review of the Work Environment Scale, Second Edition. Twelfth Mental Measurements Yearbook, 1121–1122.
MULAIK, S. A., JAMES, L. R., VAN ALSTINE, J., BENNETT, N., LIND, S., & STILWELL, C. D. (1989). Evaluation of goodness-of-fit indices for structural equation models. Psychological Bulletin, 105, 430–445.
MULCAHY, R. F., SHORT, R. H., & ANDREWS, J. (Eds.). (1991). Enhancing learning and thinking. New York: Praeger.

MULLEN, J. D., & ROTH, B. M. (1991). Decision-making: Its logic and practice. Savage, MD: Rowman & Littlefield.
MULLEN, Y. (1992). Assessment of the preschool child with hearing impairment. In E. Vazquez Nutall, I. Romero, & J. Kalesnik (Eds.), Assessing and screening preschoolers: Psychological and educational dimensions (pp. 327–343). Boston: Allyn & Bacon.
MUMFORD, M. D., & STOKES, G. S. (1992). Developmental determinants of individual action: Theory and practice in applying background measures. In M. D. Dunnette & L. M. Hough, (Eds.), Handbook of industrial and organizational psychology, (2nd ed., Vol. 3, pp. 61–138). Palo Alto, CA: Consulting Psychologist Press.
MUMFORD, M. D., STOKES, G. S., & OWENS, W. A. (1990). Patterns of life adaptation: The ecology of human individuality. Hillsdale, NJ: Erlbaum.
MURPHY, G., & KOVACH, J. R. (1972). Historical introduction to modern psychology (3rd ed.). San Diego, CA: Harcourt, Brace, Jovanovich.
MURPHY, K. R. (1992). Review of the Test of Nonverbal Intelligence, Second Edition. Eleventh Mental Measurements Yearbook, pp. 969–970.
MURPHY, K. R. (1993). Honesty in the workplace. Pacific Grove, CA: Brooks/Cole.
MURPHY, K. R., & ANHALT, R. L. (1992). Is halo error a property of the rater, ratees, or the specific behaviors observed? Journal of Applied Psychology, 77, 494–500.
MURPHY, K. R., & BALZER, W. K. (1989). Rater errors and rating accuracy. Journal of Applied Psychology, 74, 619–624.
MURRAY, H. A., et al. (1938). Explorations in personality: A clinical and experimental study of fifty men of college age. New York: Oxford University Press.
MURRAY, H. A., et al. (1943). Thematic Apperception Test: Manual. Cambridge, MA: Harvard University Press.
MURRAY, H. A., & MACKINNON, D. W. (1946). Assessment of OSS personnel. Journal of Consulting Psychology, 10, 76–80.
MURSTEIN, B. I. (1963). Theory and research in projective techniques (emphasizing the TAT). New York: Wiley.
MURSTEIN, B. I. (1972). Normative written TAT responses for a college sample. Journal of Personality Assessment, 36, 213–217.
MUSSEN, P. H., & NAYLOR, H. K. (1954). The relationships between overt and fantasy aggression. Journal of Abnormal and Social Psychology, 49, 235–240.
MYERS, H. R, WOHLFORD, P., GUZMAN, L. P., & ECHEMENDIA, R. J. (Eds.). (1991). Ethnic minority perspective on clinical training and services in psychology. Washington, DC: American Psychological Association.
MYERS, I. B. (1962). Manual: The Myers-Briggs Type Indicator. Princeton, NJ: Educational Testing Service.
MYERS, I. B., & McCAULLEY, M. H. (1985). Manual: A guide to the development and use of the Myers-Briggs Type Indicator. Palo Alto, CA: Consulting Psychologists Press.
NADIEN, M. B. (1989). Adult years and aging. Dubuque, IA: Kendall/Hunt.
NAGLIERI, J. A. (1988). Draw A Person: A quantitative scoring system—Manual. San Antonio, TX: Psychological Corporation.
NAGLIERI, J. A., & DAS, J. P. (1990). Planning, attention, simultaneous, and successive (PASS) cognitive processes as a model for intelligence. Journal of Psychoeducational Assessment, 8, 303–337.
NAGLIERI, J. A., & DAS, J. P. (1997a). Das-Naglieri Cognitive Assessment System: Administration and scoring manual. Itasca, IL: Riverside.
NAGLIERI, J. A., & DAS, J. P. (1997b). Das-Naglieri Cognitive Assessment System: Interpretive handbook. Itasca, IL: Riverside.
NAGLIERI, J. A., & PFEIFFER, S. I. (1992). Performance of disruptive behavior disordered and normal samples on the Draw A Person: Screening Procedure for Emotional Disturbance. Psychological Assessment, 4, 156–159.
NAGLIERI, J. A., & PREWETT, P. N. (1990). Nonverbal intelligence measures: A selected review of instruments and their use. In C. R. Reynolds & R. W. Kamphaus (Eds.), Handbook of psychological and educational assessment of children: Intelligence and achievement (pp. 348–370). New York: Guilford Press.
NATHAN, B. R. (1986). The halo effect: It is a unitary concept! Journal of Occupational Psychology, 59, 41–44.
NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS (NAEP). (1985). The reading report card: Progress toward excellence in our schools (NAEP Report 15-R-01). Princeton, NJ: Author.
NATIONAL COMMISSION ON TESTING AND PUBLIC POLICY. (1990). From gatekeeper to gateway: Transforming testing in America. Chestnut Hill, MA: Boston College and Author.
NATIONAL COUNCIL ON EDUCATION STANDARDS AND TESTING. (1992). Raising standards for American education: A Report to Congress, the Secretary of Education, the National Education Goals Panel, and the American people. Washington, DC: Author.
NAYLOR, J. C., & SHINE, L. C. (1965). A table for determining the increase in mean criterion score obtained by using a selection device. Journal of Industrial Psychology, 3, 33–42.
NEIMARK, E. D. (1987). Adventures in thinking. San Diego, CA: Harcourt Brace Jovanovich. (Ed.).
NEIMEYER, G. J. (1989). Applications of repertory grid technique to vocational assessment. Journal of Counseling and Development, 67, 585–589.
NEIMEYER, G. J. (Ed).) (1993). Constructivist assessment: A casebook. Thousand Oaks, CA: Sage.
NEIMEYER, G. J., & NEIMEYER, R. A. (Eds.). (1990). Advances in personal construct psychology (Vol. 1). Greenwich, CT: JAI Press.
NEIMEYER, R. A., & MAHONEY, M. J. (Eds.). (1995). Constructivism in psychotherapy. Washington, DC: American Psychological Association.
NEIMEYER, R. A., & NEIMEYER, G. J. (Eds.). (1992). Advances in personal construct psychology (Vol. 2). Greenwich, CT: JAI Press.
NEISSER, U. (1976). General, academic, and artificial intelligence. In L. B. Resnick (Ed.), The nature of intelligence (pp. 135–144). Hillsdale, NJ: Erlbaum.
NEISSER, U. (1979). The concept of intelligence. Intelligence, 3, 217–227.
NEISSER, U., BOODOO, G., BOUCHARD, T. J., JR., BOYKIN, A. W., BRODY, N., CECI, S. J., HALPERN, D. F., LOEHLIN, J. C., PERLOFF, R., STERNBERG, R. J., & URBINA, S. (1996). Intelligence: Knowns and unknowns. American Psychologist, 51, 77–101.
NELSON, R. O., & HAYES, S. C. (1986). The nature of behavioral assessment. In R. O. Nelson & S. C. Hayes (Eds.), Conceptual foundations of behavioral assessment (pp. 3–41). New York: Guilford Press.

NESSELROADE, J. R., & REESE, H. W. (Eds.). (1973). Life-span developmental psychology: Methodological issues. New York: Academic Press.
NESSELROADE, J. R., & VON EYE, A. (Eds.). (1985). Individual development and social change: Exploratory analysis. Orlando, EL: Academic Press.
NESTER, M. A. (1994). Psychometric testing and reasonable accommodation for persons with disabilities. In S. M. Bruyere & J. O'Keeffe (Eds.), Implications of the Americans with Disabilities Act for psychology (pp. 25–36). Washington, DC: American Psychological Association.
NETTER, B. E. C., & VIGLIONE, D. J., JR. (1994). An empirical study of malingering schizophrenia on the Rorschach. Journal of Personality Assessment, 62, 45–57.
NEUFELDT, S. A., IVERSEN, J. N., & JUNTUNEN, C. L. (1995). Supervision strategies for the first practicum. Alexandria, VA: American Counseling Association,
NEVILL, D. D., & SUPER, D. E. (1989). The Values Scale: Theory, application, and research — Manual (2nd ed.). Palo Alto, CA: Consulting Psychologists Press.
NEVO, B. (1985). Face validity revisited. Journal of Educational Measurement, 22, 287–293.
NEVO, B. (1992). Examinee feedback: Practical guidelines. In M. Zeidner & R. Most (Eds.), Psychological testing: An inside view (pp. 377–398). Palo Alto, CA: Consulting Psychologists Press.
NEVO, B., & JAGER, R. S. (Eds.). (1993). Educational and psychological testing: The test taker's outlook. Gottingen, Germany: Hogrefe & Huber.
NEVO, B., & SFEZ, J. (1985). Examinees' feedback questionnaires: Assessment and Evaluation in Higher Education, 10, 236–249.
NEVO, O., & NEVO, B. (1983). What do you do when asked to answer humorously? Journal of Personality and Social Psychology, 44, 188–194.
NEWELL, A., & SIMON, H. A. (1972). Human problem solving. Englewood Cliffs, NJ: Prentice Hall.
NEWLAND, T. E. (1979). The Blind Learning Aptitude Test. Journal of Visual Impairment and Blindness, 73,134–139.
NIAZ, M. (1987). Mobility-fixity dimension in Witkin's theory of field-dependence/ independence and its implications for problem solving in science. Perceptual and Motor Skills, 65, 755–764.
NICHOLS, D. S. (1992). Review of the Minnesota Multiphasic Personality Inventory–2. Eleventh Mental Measurements Yearbook, 562–565.
NICHOLS, D. S., & GREENE, R. L. (1995). MMPI–2 structural summary: Interpretive manual. Odessa, FL: Psychological Assessment Resources.
NICHOLS, J. G. (1979). Quality and equality in intellectual development: The role of motivation in education. American Psychologist, 34, 1071–1084.
NICHOLS, P. L., & BROMAN, S. H. (1974). Familial resemblance in infant mental development. Developmental Psychology, 10, 442–446.
NICHOLSON, C. L., & ALCORN, C. L. (1994). Educational applications of the WISC-III: A handbook of interpretive strategies and remedial recommendations. Los Angeles: Western Psychological Services.
NICKERSON, R. S. (1988). On improving thinking through instruction. Review of Research in Education, 15, 3–57.
NIHIRA, K., LELAND, H., & LAMBERT, N. (1993). AAMR Adaptive Scale—Residential and Community—Second Edition: Examiner s Manual. Austin, TX: PRO-ED.
The Ninth Mental Measurements Yearbook. (1985). Lincoln, NE: Buros Institute of Mental Measurements.
NISBET, J. D. (1957). Symposium: Contributions to intelligence testing and the theory of intelligence: IV. Intelligence and age: Retesting with twenty-four years' interval. British Journal of Educational Psychology, 27, 190–198.
NITKO, A. J. (1984). Defining «criterion-referenced test.» In R. A. Berk (Ed.), A guide to criterion-referenced test construction (pp. 8–28). Baltimore: Johns Hopkins University Press.
NITKO, A. J. (1989). Designing tests that are integrated with instruction. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 447–474). New York: American Council on Education/Macmillan.
NORRIS, L., SCHOTT, P. S., SHATKIN, L., & BENNETT, M. F. (1986). The development and field testing of SIGI PLUS (ETS Res. Mem., 86–6). Princeton, NJ: Educational Testing Service.
NOVICK, M. R., & LEWIS, C. (1967). Coefficient alpha and the reliability of composite measurements. Psychometrika, 32, 1–13.
NOVY, D. M. (1992). Gender comparability of Forms 81 of the Washington University Sentence Completion Test. Educational and Psychological Measurement, 52, 491–497.
NOVY, D. M., & FRANCIS, D. J. (1992). Psychometric properties of the Washington University Sentence Completion Test. Educational and Psychological Measurement, 52, 1029–1039.
NOVY, D. M., GAA, J. P., FRANKIEWICZ, R. G., LIBERMAN, D., & AMERIKANER, M. (1992). The association between patterns of family functioning and ego development of the juvenile offender. Adolescence, 27, 25–35.
NOWICKI, S., JR., & DUKE, M. P. (1983). The Nowicki-Strickland life-span locus of control scales: Construct validation. In H. M. Lefcourt (Ed.), Research with the locus of control construct (Vol. 2, pp. 13–51). Orlando, FL: Academic Press.
NUGENT, J. K., LESTER, B. M., & BRAZELTON, T. B. (Eds.). (1991). The cultural context of infancy, Vol. 2: Multicultural and interdisciplinary approaches to parent-infant relations. Norwood, NJ: Ablex.
OAKLAND, T., GLUTTING, J., & HORTON, C. (1996). Student Styles Questionnaire: Manual. San Antonio, TX: Psychological Corporation.
OAKLAND, T., & HAMBLETON, R. K. (Eds.). (1995). International perspectives on academic assessment. Boston: Kluwer.
OAKLAND, T., & HU, S. (1992). The top 10 tests used with children and youth world wide. Bulletin of the International Test Commission, 19, 99–120.
O'BRIEN, W. H., & HAYNES, S. N. (1993). Behavioral assessment in the psychiatric setting. In A. S. Bellack & M. Hersen (Eds.), Handbook of behavior therapy in the psychiatric setting (pp. 39–71). New York: Plenum Press.
OBRZUT, J. E., & BOLIEK, C. A. (1986). Thematic approaches to personality assessment with children and adolescents. In H. M. Knoff (Ed.), The assessment of child and adolescent personality (pp. 173–198). New York: Guilford Press.

GETTING, E. R., & DEFFENBACHER, J. L. (1980). Text Anxiety Profile manual. Fort Collins, CO: Rocky Mountain Behavioral Science Institute.
OFFICE OF TECHNOLOGY ASSESSMENT. (1992). Testing in American schools: Asking the right questions (QTA-SET-520). Washington, DC: U. S. Government Printing Office.
OGILVIE, D. M., & ASHMORE, R. D. (1991). Self-with-other representation as a unit of analysis in self-concept research. In R. C. Curtis (Ed.), The relational self (pp. 282–314). New York: Guilford Press.
OLES, H. J., & DAVIS, G. D. (1977). Publishers violate APA standards on test distribution. Psychological Reports, 41, 713–714.
OLKIN, I., & FINN, J. D. (1995). Correlations redux. Psychological Bulletin, 128, 155–164.
OLLENDICK, T. H., & HERSEN, M. (Eds.). (1993). Handbook of child and adolescent assessment. Boston: Allyn & Bacon.
OLSAT, 7th ed.: Technical manual. (1997). San Antonio, TX: Harcourt Brace.
OLSON, J. M., & ZANNA, M. P. (1993). Attitudes and attitude change. Annual Review of Psychology, 44,117–154.
OLSON-BUCHANAN, J. B., DRASGOW, F., MOBERG, P. J., MEAD, A. D., & KEENAN, P. A. (1996). The Conflict Resolution Skills Assessment: Model-based, multi-media measurement. Manuscript submitted for publication.
OLTON, R. M., & CRUTCHFIELD, R. S. (1969). Developing the skills of productive thinking. In P. H. Mussen, J. Langer, & M. Covington (Eds.), Trends and issues in developmental psychology (pp. 68–91). New York: Holt, Rinehart & Winston.
On your own: Preparing for a standardized test (videodisk). (1987). Princeton, NJ: Educational Testing Service.
ONES, D. S., VISWESVARAN, C., & SCHMIDT, F. L. (1993). Comprehensive meta-analysis of integrity test validities: Findings and implications for personnel selection and theories of job performance. Journal of Applied Psychology Monographs, 78, 679–703.
OOSTERHOF, A. C. (1976). Similarity of various item discrimination indices. Journal of Educational Measurement, 23, 145–150.
OOSTERVELD, P. (1994). Confirmatory factor analysis of the Self-Directed Search test: A multitrait-multimethod approach. Personality and Individual Differences, 17, 565–569.
OOSTERWEGEL, A., & OPPENHEIMER, L. (1993). The self-system: developmental changes between and within self-concepts. Hillsdale, NJ: Erlbaum.
ORLANSKY, M. D. (1988). Assessment of visually impaired infants and preschool children. In T. D. Wachs & R. Sheehan (Eds.), Assessment of young developmentally disabled children (pp. 93–107). New York: Plenum Press.
ORTAR, G. (1963). Is a verbal test cross-cultural? Scripta Hierosolymitana, 13, 219–235.
ORTAR, G. (1972). Some principles for adaptation of psychological tests. In L. J. Cronbach & P. J. D. Drenth (Eds.), Mental tests and cultural adaptation (pp. 111–120). The Hague: Mouton.
OSGOOD, C. E., SUCI, G. J., & TANNENBAUM, P. H. (1957). The measurement of meaning. Urbana: University of Illinois Press.
OSIPOW, S. H. (1973). Theories of career development (2nd ed.). New York: Appleton-Century-Crofts.
OSS ASSESSMENT STAFF. (1948). Assessment of men: Selection of personnel for the Office of Strategic Services. New York: Rinehart.
OSTERLIND, S. J. (1983). Test item bias. Newbury Park, CA: Sage.
OSTROM, T. M., BOND, C. F., JR., KROSNICK, J. A., & SEDIKIDES, C. (1994). Attitude scales: How we measure the unmeasureable. In S. Shavitt & T. C. Brock (Eds.), Persuasion: Psychological insights and perspectives (pp. 15–42). Boston: Allyn & Bacon.
OWENS, W. A. (1953). Age and mental abilities: A longitudinal study. Genetic Psychology Monographs, 48, 3–54.
OWENS, W A. (1966). Age and mental abilities: A second adult follow-up. Journal of Educational Psychology, 57, 311–325.
OWENS, W. A. (1983). Background data. In M. D. Dunnette (Ed.), Handbook of industrial and organizational psychology (pp. 609–644). New York: Wiley.
OWENS, W. A., & SCHOENFELDT, L. F. (1979). Toward a classification of persons [Monograph]. Journal of Applied Psychology, 64, 569–607.
OWINGS, R. A., PETERSEN, G. A., BRANSFORD, J. D., MORRIS, C. D., & STEIN, B. S. (1980). Spontaneous monitoring and regulation of learning: A comparison of successful and less successful fifth graders. Journal of Educational Psychology, 72, 250–256.
OWNBY, R. L. (1991). Psychological reports: A guide to report writing in professional psychology (2nd ed.). Brandon, VT: Clinical Psychology Publishing Co.
OZER, D. J. (1993). The Q-sort method and the study of personality development. In D. C. Funder, R. D. Parke, C. Tomlinson-Keasey, & K. Widaman (Eds.), Studying lives through time: Personality and development (pp. 147–168). Washington, DC: American Psychological Association.
OZER, D. J., & REISE, S. P. (1994). Personality assessment. Annual Review of Psychology, 45, 357–388.
PAAJANEN, G. E., HANSEN, T. L., & McLELLAN, R. A. (1993). PDI Employment Inventory and PDI Customer Service Inventory manual. Minneapolis, MN: Personnel Decisions.
PAGE, E. B. (1985). Review of Kaufman Assessment Battery for Children. Ninth Mental Measurements Yearbook, Vol. 1, 773–777.
PAGET, K. D. (1991). Fundamentals of family assessment. In B. A. Bracken (Ed.), The psychoeducational assessment of preschool children (2nd ed., pp. 514–528). Boston: Allyn & Bacon.
PALERMO, D. S., & JENKINS, J. J. (1963). Frequency of superordinate responses to a word association test as a function of age. Journal of Verbal Learning and Verbal Behavior, 1, 378–383.
PALISIN, H. (1986). Preschool temperament and performance on achievement tests. Developmental Psychology, 22, 766–770.
PALMORE, E. (Ed.). (1970). Normal aging. Durham, NC: Duke University Press.
PALOMARES, R. S., CROWLEY, S. L., WORCHEL, F. R, OLSON, T. K., & RAE, W. A. (1991). The factor analytic structure of the Roberts Apperception Test for Children: A comparison of the standardization sample with a sample of chronically ill children. Journal of Personality Assessment, 56,414–425.
PANELL, R. C., & LAABS, G. J. (1979). Construction of a criterion-referenced, diagnostic test for an individualized instruction program. Journal of Applied Psychology, 64, 255–261.
PANIAGUA, F. A. (1994). Assessing and treating culturally diverse clients: A practical guide. Thousand Oaks, CA: Sage.

PARKER, K. C. H., HANSON, R. K., & HUNSLEY, J. (1988). MMPI, Rorschach, and WAIS: A meta-analytic comparison of reliability, stability, and validity. Psychological Bulletin, 103, 367–373.
PARKER, R. M. (199 la). Occupational Aptitude Survey and Interest Schedule, Second Edition (OASIS–2) — Aptitude Survey: Examiners manual. Austin, TX: PRO-ED.
PARKER, R. M. (1991b). Occupational Aptitude Survey and Interest Schedule, Second Edition (OASIS–2) — Interest Schedule: Examiner's manual. Austin, TX: PRO-ED.
PARKERSON, J. A., LOMAX, R. G., SCHILLER, D. P., & WALBERG, H. J. (1984). Exploring causal models of educational achievement. Journal of Educational Psychology, 76, 638–646.
PASCAL, G. R., & SUTTELL, B. J. (1951). The Bender-Gestalt Test: Quantification and validity for adults. New York: Grune & Stratton.
PASCUAL-LEONE, J., & IJAZ, H. (1991). Mental capacity testing as a form of intellectual-developmental assessment. In R. J. Samuda, S. L. Kong, J. Cummins, J. Pascual-Leone, & J. Lewis (Eds.), Assessment and placement of minority students (pp. 143–171). Toronto: Hogrefe.
PASHLEY, P. J. (1992). Graphical IRT-based DIF analyses (Res. Rep. No. 92–66). Princeton, NJ: Educational Testing Service.
PAUL, G. L., & ERIKSEN, C. W. (1964). Effects of test anxiety on «real-life» examinations. Journal of Personality, 32, 480–494.
PAULHUS, D. L. (1983). Sphere-specific measures of perceived control. Journal of Personality and Social Psychology, 44, 1253–1265.
PAULHLJS, D. L. (1984). Two-component models of socially desirable responding. Journal of Personality and Social Psychology, 46, 598–609.
PAULHUS, D. L. (1986). Self-deception and impression management in test responses. In A. Angleitner & J. S. Wiggins (Eds.), Personality assessment via questionnaires: Current issues in theory and measurement (pp. 143–165). Berlin: Springer-Verlag.
PAULHUS, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), Measures of personality and social psychological attitudes. San Diego, CA: Academic Press.
PAULHUS, D. L., & BRUCE, M. N. (1992). The effect of acquaintanceship on the validity of personality impressions: A longitudinal study. Journal of Personality and Social Psychology, 63, 816–824.
PAULHUS, D. L., & REID, D. B. (1991). Enhancement and denial in socially desirable responding. Journal of Personality and Social Psychology, 60, 307–317.
PAUNONEN, S. V. (1993, August). Sense, nonsense, and the Big Five Factors of Personality. Paper presented at the convention of the American Psychological Association, Toronto, Canada.
PAUNONEN, S. V., JACKSON, D. N., TRZEBINSKI, J., & FORSTERLING, F. (1992). Personality structure across cultures: A multimethod evaluation. Journal of Personality and Social Psychology, 62, 447–456.
PAYNE, R. N. (1985). Review of the SCL–90-R. Ninth Mental Measurements Yearbook, Vol. 2,1326–1329.
PEARLMAN, K., SCHMIDT, F. L., & HUNTER, J. E. (1980). Validity generalization results for tests used to predict job proficiency and training success in clerical occupations. Journal of Applied Psychology, 65, 373–406.
PEARSON, K. (1901). On lines and planes of closest fit to systems of points in space. Philosophical Magazine (Series 6), 2, 559–572.
PEDERSEN, P. B. (1987). Handbook of cross-cultural counseling and therapy. Westport, CT: Greenwood.
PEDERSEN, P. B., & IVEY, A. (1993). Culture-centered counseling and interviewing skills: A practical guide. Westport, CT: Greenwood.
PEEL, E. A. (1951). A note on practice effects in intelligence tests. British Journal of Educational Psychology, 21, 122–125.
PEEL, E. A. (1952). Practice effects between three consecutive tests of intelligence. British Journal of Educational Psychology, 22, 196–199.
PELLEGRINQ, J. W., & GLASER, R. (1979). Cognitive correlates and components in the analysis of individual differences. Intelligence, 3, 187–214.
PELLEGRINO, J. W., MUMAW, R. J., & SHUTE, V. J. (1985). Analyses of spatial aptitude and expertise. In S. E. Embretson (Ed.), Test design: Developments in psychology and psychometrics (pp. 45–76). Orlando, FL: Academic Press.
PEMBERTON, C. L. (1952). The closure factors related to temperament. Journal of Personality.21, 159–175.
PENNER, L. A., BATSCHE, G. M., KNOFF, H. M., & NELSON, D. L. (Eds.). (1993). The challenge in mathematics and science education: Psychology's response. Washington, DC: American Psychological Association.
PENNINGTON, B. F. (1991). Diagnosing learning disorders: A neuropsychological framework. New York: Guilford Press.
PENNOCK-ROMAN, M. (1990). Test validity and language background: A study of Hispanic-American students at six universities. New York: College Entrance Examination Board.
PERRY, G. G., & KINDER, B. N. (1990). The susceptibility of the Rorschach to malingering: A critical review. Journal of Personality Assessment, 54, 47–57.
PERRY, W. (1993). Rorschach for the '90s: An interpretation milestone. Journal of Personality Assessment, 60, 418–420.
PETERSEN, N. S., KOLEN, M. J., & HOOVER, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 221–262). New York: American Council on Education/Macmillan.
PETERSEN, N. S., & NOVICK, M. R. (1976). An evaluation of some models for culture-fair selection. Journal of Educational Measurement, 13, 3–29.
PETERSON, C. A. (1994). Book review: The Eleventh Mental Measurements Yearbook. Journal of Personality Assessment, 63, 394–397.
PETERSON, D. (1968). The clinical study of social behavior. New York: Appleton-Century-Crofts.
PETERSON, G. W., SAMPSON, J. P., JR., & REARDON, R. C. (1991). Career development and services: A cognitive approach. Pacific Grove, CA: Brooks/Cole.
PETERSON, J. (1926). Early conceptions and tests of intelligence. Yonkers, NY: World Book.
PETERSON, N. G., HOUGH, L. M., DUNNETTE, M. D., ROSSE, R. L., HOUSTON, J. S., TOQUAM, J. L., & WING, H. (1990). Project A: Specification of the predictor domain and development of new selection/classification tests. Personnel Psychology, 43, 247–276.

PETRILA, J. & OTTO, R. K. (1995). Law and mental health professionals: Florida. Washington, DC: American Psychological Association.
PHILIPPE, J. (1894). Jastrow — exposition d'anthropologie de Chicago — testes psychologiques, etc. Annee Psychologique, 1, 522–526.
PIACENTINI, J. (1993). Checklists and rating scales. In T. H. Ollendick & M. Hersen (Eds.). Handbook of child and adolescent assessment (pp. 82–97). Boston: Allyn & Bacon.
PIAGET, J. (1972). Intellectual evolution from adolescence to adulthood. Human Development, 15, 1–12.
PICKMAN, A. J. (1994). The complete guide to outplacement counseling. Hillsdale, NJ: Erlbaum.
PIEDMONT, R. L., McCRAE, R. R., & COSTA, P. T, JR. (1992). An assessment of the Edwards Personal Preference Schedule from the perspective of the Five-Factor Model. Journal of Personality Assessment, 58, 67–78.
PIETROFESA, J. J., & SPLETE, H. (1975). Career development: Theory and research. Orlando, FL: Grune & Stratton.
PINARD, A., & LAURENDEAU, M. (1964). A scale of mental development based on the theory of Piaget: Description of a project. Journal of Research in Science Teaching, 2, 253–260.
PINDER, C. C. (1973). Statistical accuracy and practical utility in the use of moderator variables. Journal of Applied Psychology, 57, 214–221.
PINNEAU, S. R. (1961). Changes in intelligence quotient from infancy to maturity. Boston: Houghton Mifflin.
PIOTROWSKI, C. (1984). The status of projective techniques: Or, «Wishing won't make it go away.» Journal of Clinical Psychology, 40, 1495–1502.
PIOTROWSKI, C., & KELLER, J. W. (1992). Psychological testing in applied settings: A literature review from 1982–1992. Journal of Training & Practice in Professional Psychology, 6, 74–82.
PIOTROWSKI, C., SHERRY, D., & KELLER, J. W. (1985). Psychodiagnostic test usage: A survey of the Society for Personality Assessment. Journal of Personality Assessment, 49, 115–119.
PIOTROWSKI, C., & ZALEWSKI, C. (1993). Training in psychodiagnostic testing in APA-approved PsyD and PhD clinical psychology programs. Journal of Personality Assessment, 61, 393–405.
PLAKE, B. S. (1980). A comparison of a statistical and a subjective procedure to ascertain item validity: One step in the test validation process. Educational and Psychological Measurement, 40, 397–404.
PLANT, W. T., & MINIUM, E. W. (1967). Differential personality development in young adults of markedly different aptitude levels. Journal of Educational Psychology, 58, 141–152.
PLOMIN, R., DEFRIES, J. C., & FULKER, D. W. (1988). Nature and nurture during infancy and early childhood. New York: Cambridge University Press.
PLOMIN, R., & McCLEARN, G. E. (Eds.). (1993). Nature, nurture, and psychology. Washington, DC: American Psychological Association.
PLOMIN, R., & READE, R. (1991). Human behavioral genetics. Annual Review of Psychology, 42, 161–190.
POON, L. W. (Ed.). (1986). Handbook for clinical memory assessment of older adults. Washington, DC: American Psychological Association.
POPE, K. S. (1992). Responsibilities in providing psychological test feedback to clients. Psychological Assessment, 4, 268–271.
POPE, K. S., BUTCHER, J. N., & SEELEN, J. (1993). The MMPI, MMPI-2, and MMPI-A in court: A practical guide for expert witnesses and attorneys. Washington, DC: American Psychological Association.
POPE, K. S., & VASQUEZ, M. J.T. (1991). Ethics in psychotherapy and counseling: A practical guide for psychologists. San Francisco: Jossey-Bass.
POPE, M. (1995). Review of the Kuder General Interest Survey, Form E. Twelfth Mental Measurements Yearbook, 543–545.
POPHAM, W. J. (1984). Specifying the domain of content or behaviors. In R. A. Berk (Ed.), A guide to criterion-referenced test construction (pp. 29–48). Baltimore: Johns Hopkins University Press.
POPHAM, W. J., & HUSEK, T. R. (1969). Implications of criterion referenced measurement. Journal of Educational Measurement, 6, 1–9.
PORTEGAL, M. (Ed.). (1982). Spatial abilities: Developmental and physiological foundations. Orlando, FL: Academic Press.
PORTEUS, S. D. (1931). The psychology of a primitive people. New York: Longmans, Green.
POSTMAN, L., & KEPPEL, G. (1970). Norms of word association. New York: Academic Press.
POTH, R. L., & BARNETT, D. W. (1988). Establishing the limits of interpretive confidence: A validity study of two preschool developmental scales. School Psychology Review, 17, 322–330.
POWELL, D. H., KAPLAN, E. F., WHITLA, D., WEINTRAUB, S., CATLIN, R., & FUNKENSTEIN, H. H. (1993). MicroCog Assessment of Cognitive Functioning: Manual. San Antonio, TX: Psychological Corporation.
POWELL, D. H., & WHITLA, D. K. (1994a). Normal cognitive aging: Toward empirical perspectives. Current Directions in Psychological Science, 3, 27–31.
POWELL, D. H., & WHITLA, D. K. (1994b). Profiles in cognitive aging. Cambridge, MA: Harvard University Press.
POWERS, D. E. (1983). Effects of coaching on GRE Aptitude Test scores (GRE Board Res. Rep. GREB No. 81–3R). Princeton, NJ: Educational Testing Service.
POWERS, D. E. (1986). Relations of test item characteristics to test preparation/test practice effects: A quantitative summary. Psychological Bulletin, 100, 67–77.
POWERS, D. E., & SWINTON, S. S. (1984). Effects of self-study for coachable test item types. Journal of Educational Psychology, 76, 266–278.
PREDIGER, D. J. (1982). Dimensions underlying Holland's hexagon: Missing link between interests and occupations? Journal of Vocational Behavior, 21, 259–287.
PREDIGER, D. J. (1993). Multicultural assessment standards: A compilation for counselors. Alexandria, VA: American Counseling Association.
PREDIGER, D. (1996). Alternative dimensions for the Tracey-Rounds interest sphere. Journal of Vocational Behavior, 48, 59–67.
PREDIGER, D. J., & VANSICKLE, T. R. (1992). Locating occupations on Holland's hexagon: Beyond RIASEC. Journal of Vocational Behavior, 40, 111–128.

PRIMOFF, E. S. (1959). Empirical validations of the J-coefficient. Personnel Psychology, 12, 413–418.
PRIMOFF, E. S. (1975). How to prepare and conduct job element examinations. Washington, DC: U.S. Government Printing Office.
PRIMOFF, E. S., & EYDE, L. D. (1988). Job element analysis. In S. Gael (Ed.), The job analysis handbook for business, industry, and government (Vol. 2, pp. 807–824). New York: Wiley.
PRINCE, J. P. (1995). Strong Interest Inventory resource: Strategies for group and individual interpretations in college settings. Palo Alto, CA: Consulting Psychologists Press.
Privacy and behavioral research,.(1967). Washington, DC: U. S. Government Printing Office.
PROCTER, M. (1993). Measuring attitudes. In N. Gilbert (Ed.), Researching social life (pp. 116–134). London: Sage.
PROVENCE, S., ERIKSON, J., VATER, S., & PALMERI, S. (1995a). Infant-Toddler Developmental Assessment – Family centered assessment of young children at risk: The IDA readings. Chicago: Riverside.
PROVENCE, S., ERIKSON, J., VATER, S., & PALMERI, S. (1995b). Infant-Toddler Developmental Assessment: Foundations and study guide. Chicago: Riverside.
PROVENCE, S., ERIKSON, J., VATER, S., & PALMERI, S. (1995c). Infant-Toddler Developmental Assessment – IDA administration manual: Procedures summary – Provence Birth-to-Three Developmental Profile. Chicago: Riverside.
PSYCHOLOGICAL CORPORATION. (1991a). Counselors Manual for Interpreting the Career Interest Inventory. San Antonio, TX: Author.
PSYCHOLOGICAL CORPORATION. (1991b). Differential Aptitude Tests, Fifth Edition/Career Interest Inventory: Counselor's manual. San Antonio, TX: Author.
PSYCHOLOGICAL CORPORATION. (1992a). Differential Aptitude Tests, Fifth Edition: Technical manual. San Antonio, TX: Author.
PSYCHOLOGICAL CORPORATION. (1992b). Wechsler Individual Achievement Test – WIAT: Manual. San Antonio, TX: Author.
PULAKOS, E. D. (1986). The development of training programs to increase accuracy with different rating tasks. Organizational Behavior and Human Decision Processes, 38, 76–91.
QUAN, B., PARK, T. A., SANDAHL, G., & WOLFE, J. H. (1984). Microcomputer network for computerized adaptive testing CAT *(Tech. Rep. 84–33). San Diego, CA: Navy Personnel Research and Development Center.*
RABIN, A. I. (Ed.). (1981). Assessment with protective techniques: A concise introduction. New York: Springer.
RABIN, A. I. (Ed.). (1986). Projective techniques for adolescents and children. New York: Springer.
RABIN, A. I., & GUERTIN, W. H. (1951). Research with the Wechsler-Bellevue Test: 1945–1950. Psychological Bulletin, 48, 211–248.
RABIN, A. I., & ZLOTOGORSKI, Z. (1981). Completion methods: Word association, sentence, and story completion. In A. I. Rabin (Ed.), Assessment with projective techniques: A concise introduction (pp. 121–149). New York: Springer.
RADCLIFFE, J. A. (1966). A note on questionnaire faking with the 16PFQ and MPI. Australian Journal of Psychology, 18, 154–157.
RAGGIO, D. J., & MASSINGALE, T. W. (1990). Comparability of the Vineland Social Maturity Scale and the Vineland Adaptive Behavior Scale—Survey form with infants evaluated for developmental delay. Perceptual and Motor Skills, 71, 415–418.
RAJU, N. S., BURKE, M. J., & NORMAND, J. (1990). A new approach to utility analysis. Journal of Applied Psychology, 75, 3–12.
RAMSEYER, G. C., & CASHEN, V. M. (1971). The effect of practice sessions on the use of separate answer sheets by first and second graders. Journal of Educational Measurement, 8,177–181.
RAND, Y., TANNENBAUM, A. J., & FEUERSTEIN, R. (1979). Effects of instrumental enrichment on the psychoeducational development of low-functioning adolescents. Journal of Educational Psychology, 71, 751–763.
RANDAHL, G. J., HANSEN, J. C., & HAVERKAMP, B. E. (1993). Instrumental behaviors. following test administration and interpretation: Exploration validity of the Strong Interest Inventory. Journal of Counseling and Development, 71, 435–439.
RAPAPORT, D., et al. (1968). Diagnostic psychological testing (rev. ed. edited by R. R. Holt). New York: International Universities Press. (Original work published 1946)
RASCH, G. (1966). An individualistic approach to item analysis. In R F. Lazarsfeld & N. W. Henry (Eds.), Readings in mathematical social sciences (pp. 89–107). Cambridge, MA: MIT Press.
RASKIN, E. (1985). Counseling implications of field dependence-independence in an educational setting. In M. Bertini, L. Pizzamiglio, & S. Wapner (Eds.), Field dependence in psychological theory, research, and application: Two symposia in memory of Herman A. Witkin (pp. 107–113). Hillsdale, NJ: Erlbaum.
RAVEN, J. (1983). The Progressive Matrices and Mill Hill Vocabulary Scale in Western Societies. In S. H. Irvine & J. W. Berry (Eds.), Human assessment and cultural factors (pp. 107–114). New York: Plenum Press.
RAVEN, J., RAVEN, J. C., & COURT, J. H. (1995). Manual for Ravens Progressive Matrices and vocabulary scales – Section 1: General Overview (1995 Edition). Oxford, England: Oxford Psychologists Press.
RECKASE, M. D. (1990). Scaling techniques. In G. Goldstein & M. Hersen (Eds.), Handbook of psychological assessment (2nd ed., pp. 41–56). Elmsford, NY: Pergamon Press.
REED, R., ROTATORI, A. F., & DAY, G. R (1990). Career and vocational assessment. In A. R Rotatori, R. A. Fox, D. Sexton, & J. Miller (Eds.), Comprehensive assessment in special education: Approaches, procedures, and concerns (pp. 341–386). Springfield, IL: Charles C Thomas.
REESE, H. W. (Ed.). (1987). Advances in child development and behavior (Vol. 20). Orlando, FL: Academic Press.
REEVES, D., & WEDDING, D. (1994). The clinical assessment of memory: A practical guide. New York: Springer.
REICHENBERG-HACKETT, W. (1953). Changes in Goodenough drawings after a gratifying experience. American Journal of Orthopsychiatry ,23,501–517.
REILLY, R. R. (1973). A note on minority group test bias studies. Psychological Bulletin, 80, 130–132.
REINEHR, R. C. (1992). Review of Differential Ability Scales. Eleventh Mental Measurements Yearbook, 282–283.
REINERT, G. (1970). Comparative factor analytic studies of intelligence throughout the human life-span. In L. R. Goulet & P. B. Baltes (Eds.), Life-span developmental psychology: Research and theory (pp. 467–484). New York: Academic Press.
REISE, S. P., & OLIVER, C. J. (1994). Development of a California Q-set indicator of primary psychopathy. Journal of Personality Assessment, 62, 130–144.

REISS, S. (1994). Issues in defining mental retardation. American Journal on Mental Retardation, 99, 1–7.
REITAN, R. M. (1955). Certain differential effects of left and right cerebral lesions in human adults. Journal of Comparative and Physiological Psychology, 48, 474–477.
REITAN, R. M. (1966). A research program on the psychological effects of brain lesions in human beings. In N. R. Ellis (Ed.), International review of research in mental retardation (Vol. 1, pp. 153–218). Orlando, FL: Academic Press.
REITAN, R. M., & WOLFSON, D. (1993). The Halstead-Reitan Neuropsychological Test Battery: Theory and clinical interpretation (2nd ed.). Tucson, AZ: Neuropsychology Press.
RENNINGER, K. A., HIDI, S., & KRAPP, A. (Eds.). (1992). The role of interest in learning and development. Hillsdale, NJ: Erlbaum.
RENTZ, R. R., & BASHAW, W. L. (1977). The National Reference Scale for reading: An application of the Rasch model. Journal of Educational Measurement, 14, 161–179.
REPP, A. C., & FELCE, D. (1990). A microcomputer system used for evaluative and experimental behavioural research in mental handicap. Mental Handicap Research, 3, 21–32.
RESCHLY, D. J. (1988). Larry P.! Larry P.! Why the California sky fell on IQ testing. Journal of School Psychology, 26, 199–205.
RESNICK, L. B. (Ed.). (1976). The nature of intelligence. Hillsdale, NJ: Erlbaum.
RESNICK, L. B., & GLASER, R. (1976). Problem solving and intelligence. In L. B. Resnick (Ed.), The nature of intelligence (pp. 205–230). Hillsdale, NJ: Erlbaum.
RESNICK, L. B., & NECHES, R. (1984). Factors affecting individual differences in learning ability. In R. J. Sternberg (Ed.), Handbook of human intelligence (Vol. 2, pp. 275–323). Hillsdale, NJ: Erlbaum.
RESNICK, L. B., & RESNICK, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. R. Gifford & M. C. O'Connor (Eds.), Changing assessments: Alternative views of aptitude, achievement and instruction (pp. 37–75). Boston: Kluwer.
RETZLAFF, P. (1992). Review of the State-Trait Anger Expression Inventory, Research Edition. Eleventh Mental Measurements Yearbook, 869–870.
RETZLAFF, P. (1995). Tactical psychotherapy of the personality disorders: An MCMI-III-based approach. Boston: Allyn & Bacon.
REYNOLDS, C. R. (1982). Methods for detecting construct and predictive bias. In.R. A. Berk (Ed.), Handbook of methods for detecting test bias (pp. 199–227). Baltimore: Johns Hopkins University Press.
REYNOLDS, C. R. (1986). Vineland Adaptive Behavior Scales, 1984 Edition. Journal of Educational Measurement, 23, 389–391.
REYNOLDS, C. R. (1990). Conceptual and technical problems in learning disability diagnosis. In C. R. Reynolds & R. W. Kamphaus (Eds.), Handbook of psychological and educational assessment of children: Intelligence and achievement (pp. 571–593). New York: Guilford Press.
REYNOLDS, C. R. (1992a). Review of the Millon Clinical Multiaxial Inventory-II. Eleventh Mental Measurements Yearbook, 533–535.
REYNOLDS, C. R. (1992b). Two key concepts in the diagnosis of learning disabilities and the habilitation of learning. Learning Disability Quarterly, 15, 2–12.
REYNOLDS, C. R., & BROWN, R. T. (1984). Perspectives on bias in mental testing. New York: Plenum Press.
REYNOLDS, C. R. & KAMPHAUS, R. W. (Eds.). (1990a). Handbook of psychological and educational assessment of children: Intelligence and achievement. New York: Guilford Press.
REYNOLDS, C. R., & KAMPHAUS, R. W. (Eds.). (1990b). Handbook of psychological and educational assessment of children: Personality, behavior, and context. New York: Guilford Press.
REYNOLDS, C. R., & KAMPHAUS, R. W. (1992). Behavior Assessment System for Children: Manual. Circle Pines, MN: American Guidance Service.
REYNOLDS, S. B. (1989). Review of the Multidimensional Aptitude Battery. Tenth Mental Measurements Yearbook, 522–523.
REZMOVIC, E. L., & REZMOVIC, V. (1980). Empirical validation of psychological constructs: A secondary analysis. Psychological Bulletin, 87, 66–71.
RICHARDSON, J. P. E., ANGLE, R. W., HASHER, L., LOGIE, R. H., & STOLTUS, E. R. (1996). Working memory and human cognition. New York: Oxford University Press.
RITCHIE, R. J. (1994). Using the assessment center method to predict senior management potential. Consulting Psychology Journal: Practice and Research, 46, 16–23.
RITZLER, B. (1993a). Test review: TEMAS (Tell-Me-A-Story). Journal of Psychoeducational Assessment, 11, 381–389.
RITZLER, B. (1993b). Thanks for the memories! Journal of Personality Assessment, 60, 208–210.
RITZLER, B., & ALTER, B. (1986). Rorschach teaching in APA-approved clinical graduate programs: Ten years later. Journal of Personality Assessment, 50, 44–49.
Riverside 2000: Technical Summary I. (1994). Chicago, IL: Riverside.
ROBERTS, G. E. (1994). Interpretive handbook for the Roberts Apperception Test for Children. Los Angeles, CA: Western Psychological Services.
ROBINSON, C., & FIEBER, N. (1988). Cognitive assessment of motorically impaired infants and preschoolers. In T D. Wachs & R. Sheehan (Eds.), Assessment of young developmentally disabled children (pp. 127–161). New York: Plenum Press.
ROBINSON, J. P., SHAVER, P. R., & WRIGHTSMAN, L. S. (Eds.). (1991). Measures of personality and social psychological attitudes. San Diego, CA: Academic Press.
ROBINSON, S. P. (1993). The politics of multiple-choice versus free-response assessment. In R. E. Bennett & W. C. Ward (Eds.), Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment (pp. 313–323). Hillsdale, NJ: Erlbaum.
ROCK, D. A., BENNETT, R. E., & JIRELE, T. (1988). Factor structure of the Graduate Record Examination's General Test in handicapped and non-handicapped groups. Journal of Applied Psychology, 73, 382–392.
RODGER, A. G. (1936). The application of six group intelligence tests to the same children, and the effects of practice. British Journal of Educational Psychology, 6, 291–305.

ROECKER, C. E. (1995). Well stated well met [Review of the book Intelligent testing with the WISC-III]. Contemporary Psychology, 40, 659–660.
ROGERS, C. R., & DYMOND, R. F. (Eds.). (1954). Psychotherapy and personality change. Chicago: University of Chicago Press.
ROGERS, R. (1995). Diagnostic and structured interviewing: A handbook for psychologists. Odessa, FL: Psychological Assessment Resources.
ROGOFF, B. (1990). Apprenticeship in thinking: Cognitive development in social context. New York: Oxford University Press.
ROGOFF, B., & CHAVAJAY, P. (1995). What's become of research on the cultural basis of cognitive development? American Psychologist, 50, 859–877.
ROGOFF, B., & LAVE, J. (Eds.). (1984). Everyday cognition: Its development in social context. Cambridge, MA: Harvard University Press.
ROGOFF, B., & MORELLI, G. (1989). Perspectives on children's development from cultural psychology. American Psychologist, 44, 343–348.
ROGOSA, D. (1979). Causal models in longitudinal research: Rationale, formulation, and interpretation. In J. R. Nesselroade & P. B. Bakes (Eds.), Longitudinal research in the study of behavior development (pp. 263–302). New York: Academic Press.
ROGOSA, D. (1980). A critique of cross-lagged correlation. Psychological Bulletin, 88, 245–258.
ROID, G. H. (1984). Generating the test items. In R. A. Berk (Ed.), A guide to criterion-referenced test construction (pp. 49–77). Baltimore: Johns Hopkins University Press.
ROID, G. H. (1986). Computer technology in testing. In B. S. Plake & J. C. Witt (Eds.), The future of testing (pp. 29–69). Hillsdale, NJ: Erlbaum.
ROID, G. H., & GORSUCH, R. L. (1984). Development and clinical use of test-interpretive programs on microcomputers. In M. D. Schwartz (Ed.), Using computers in clinical practice (pp. 141–149). New York: Haworth.
ROID, G. H., & MILLER, L. J. (1997). Examiner's manual: Letter International Performance Scale-Revised, Wood Dale, IL: Stocking.
RONAN, G. E, COLAVITO, V. A., & HAMMONTREE, S. R. (1993). Personal problem-solving system for scoring TAT responses: Preliminary validity and reliability data. Journal of Personality Assessment, 61, 28–40.
RONAN, G. R, DATE, A. L., & WEISBROD, M. (1995). Personal problem-solving scoring of the TAT: Sensitivity to training. Journal of Personality Assessment, 64, 119–131.
RONNING, R. R., GLOVER, J. A., CONOLEY, J. C., & WITT, J. C. (Eds.). (1987). The influence of cognitive psychology on testing. Hillsdale, NJ: Erlbaum.
ROONEY, J. P. (1987). Golden Rule on «Golden Rule.» Educational Measurement: Issues & Practice, 6, 9–12.
ROPER, B. L., BEN-PORATH, Y. S., & BUTCHER, J. N. (1991). Comparability of computerized adaptive and conventional testing with the MMPI–2. Journal of Personality Assessment, 57, 278–290.
ROPER, B. L., BEN-PORATH, Y. S., & BUTCHER, J. N. (1995). Comparability and validity of computerized adaptive testing with the MMPI–2. Journal of Personality Assessment, 65, 358–371.
RORER, L. G. (1965). The great response-style myth. Psychological Bulletin, 63, 129–156.
RORER, L. G., HOFFMAN, P. J., & HSIEH, K. (1966). Utilities as base rate multipliers in the determination of optimum cutting scores for the discrimination of groups of unequal size and variance. Journal of Applied Psychology, 50, 364–368.
RORSCHACH, H. (1942). Psychodiagnostics: A diagnostic test based on perception (P. Lemkau & B. Kronenberg, Trans.). Beme: Huber. (1st German ed. published 1921; U. S. distributor, Grune & Stratton)
ROSENBERG, M. (1965). Society and the adolescent self-image, Princeton, NJ: Princeton University Press.
ROSENTHAL, A. C. (1985). Review of Assessment in infancy: Ordinal Scales of Psychological Development. Ninth Mental Measurements Yearbook, Vol. 1, 85–86.
ROSENTHAL, R. (1966). Experimenter effects in behavioral research. New York: Appleton-Century-Crofts.
ROSENTHAL, R. (1991). Meta-analytic procedures for social research (rev. ed.). Newbury Park, CA: Sage.
ROSENTHAL, R., & ROSNOW, R. L. (Eds.). (1969). Artifact in behavioral research. New York: Academic Press.
ROSENZWEIG, S. (1950). Revised scoring manual for the Rosenzweig Picture-Frustration Study, Form for Adults. St. Louis, MO: Author.
ROSENZWEIG, S. (1960). The Rosenzweig Picture-Frustration Study, Children's Form. In A. I. Rabin & M. Haworth (Eds.), Projective techniques with children (pp. 149–176). Orlando, FL: Grune & Stratton.
ROSENZWEIG, S. (1970). Sex differences in reaction to frustration among adolescents. In J. Zubin & A. M. Freedman (Eds.), Psychopathology of adolescence (pp. 90–107). Orlando, FL: Grune & Stratton.
ROSENZWEIG, S. (1976a). Aggressive behavior and the Rosenzweig Picture-Frustration (P-F) Study. Journal of Clinical Psychology, 32, 885–891.
ROSENZWEIG, S. (1976b). Manual for the Rosenzweig Picture-Frustration Study, Adolescent Form. St. Louis, MO: Author.
ROSENZWEIG, S. (1977). Manual for the Children's Form of the Rosenzweig Picture-Frustration Study. St. Louis, MO: Rana House.
ROSENZWEIG, S. (1978a). Adult Form supplement to the basic manual of the Rosenzweig Picture-Frustration (P-F) Study. St. Louis, MO: Rana House.
ROSENZWEIG, S. (1978b). Aggressive behavior and the Rosenzweig Picture-Frustration. New York: Praeger.
ROSENZWEIG, S. (1978c). An investigation of the reliability of the Rosenzweig Picture' Frustration (P-F) Study, Children's Form. Journal of Personality Assessment, 42, 483–488.
ROSENZWEIG, S. (1978d). The Rosenzweig Picture-Frustration (P-F) Study: Basic manual. St. Louis, MO: Rana House.
ROSENZWEIG, S. (1981a). Adolescent Form supplement to the basic manual of the Rosenzweig Picture-Frustration (P-F) Study. St. Louis, MO: Rana House.
ROSENZWEIG, S. (1981b). Children's Form supplement to the basic manual of the Rosenzweig Picture-Frustration (P-F) Study. St. Louis, MO: Rana House.
ROSENZWEIG, S. (1988). Revised norms for the Children's Form of the Rosenzweig Picture-Frustration (P-F) Study, with updated reference list. Journal of Clinical Child Psychology, 17, 326–328.

ROSENZWEIG, S., & ADELMAN, S. (1977). Construct validity of the Picture-Frustration Study. Journal of Personality Assessment, 41, 578–588.
ROSS, B. M. (1991). Remembering the personal past: Descriptions of autobiographical memory. New York: Oxford University Press.
ROTHSTEIN, H. R., SCHMIDT, F. L., ERWIN, F. W, OWENS, W. A., & SPARKS, C. P. (1990). Biographical data in employment selection: Can validities be made generalizable? Journal of Applied Psychology, 75, 175–184.
ROTTER, J. B. (1966). Generalized expectancies for internal versus external control of reinforcement. Psychological Monographs, 80 (1, Whole No. 609).
ROTTER, J. B., LAH, M. I., & RAFFERTY, J. E. (1992). Rotter Incomplete Sentences Blank manual. San Antonio, TX: Psychological Corporation.
ROTTER, J. B., & RAFFERTY, J. E. (1950). Manual: The Rotter Incomplete Sentences Blank. San Antonio, TX: Psychological Corporation.
ROUNDS, J. (1995). Vocational interests: Evaluating structural hypotheses. In D. Lubinski & R. V. Dawis (Eds.), Assessing individual differences in human behavior: New concepts, methods, and findings (pp. 177–232). Palo Alto, CA: Davies-Black.
ROUNDS, J., & TRACEY, T. J. (1996). Cross-cultural structural equivalence of RIASEC models and measures. Journal of Counseling Psychology, 43, 310–329.
ROURKE, B. P. (Ed.). (1990). Neuropsychological validation of learning disability subtypes. New York: Guilford Press.
ROVEE-COLLIER, C., & LIPSITT, L. P. (Eds.). (1992). Advances in infancy research (Vol. 7). Norwood, NJ: Ablex.
ROWE, D. C. (1987). Resolving the person-situation debate: Invitation to an interdisciplinary dialogue. American Psychologist, 42, 218–227.
ROWE, H. A. H. (Ed.). (1991). Intelligence: Reconceptualization and measurement. Hillsdale, NJ: Erlbaum.
RUBIN, D. C. (Ed.). (1986). Autobiographical memory. New York: Cambridge University Press.
RULON, P. J. (1939). A simplified procedure for determining the reliability of a test of split-halves. Harvard Educational Review, 9, 99–103.
RUMSEY, M. G., WALKER, C. B., & HARRIS, J. H. (Eds.). (1994). Personnel selection and classification. Hillsdale, NJ: Erlbaum.
RUNCO, M. A. (1991). Divergent thinking. Norwood, NJ: Ablex.
RUNCO, M. A. (Ed.). (1994). Problem ending, problem solving, and creativity. Norwood, NJ: Ablex.
RUNYON, R. T., & HABER, A. (1991). Fundamentals of behavioral statistics (7th ed.). New York: McGraw'Hill.
RUSHTON, J. P. (1984). The altruistic personality: Evidence from laboratory, naturalistic, and self-report perspectives. In E. Straub, D. Bar-Tal, J. Karylowski, & J. Reykowski (Eds.), Development and maintenance of prosocial behavior (pp. 271–290). New York: Plenum Press.
RUSSELL, C. J., MATTSON, J., DEVLIN, S. E., & ATWATER, D. (1990). Predictive validity of biodata items generated from retrospective life experience essays. Journal of Applied Psychology, 75, 569–580.
RUSSELL, E. W., & STARKEY, R. I. (1993). Halstead Russell Neuropsychological Evaluation System (HRNES): Manual. Los Angeles: Western Psychological Services.
RUSSELL, M. T., & KAROL, D. (1994). Administrator's manual for the 16PF Fifth Edition. Champaign, IL: Institute for Personality and Ability Testing.
RUSSELL, T. L., REYNOLDS, D. H., & CAMPBELL, J. P. (1994). Building a joint-service classification research roadmap: Individual differences measurement (AL/HR-TP–1994-0009). Brooks AFB, TX: Armstrong Laboratory.
RUTTER, M., & RUTTER, M. (1993). Developing minds: Challenge and continuity across the life span. New York: Basic Books.
RYAN, J. J., & BOHAC, D. L. (1994). Neurodiagnostic implications of unique profiles of the Wechsler Adult Intelligence Scale-Revised. Psychological Assessment, 6, 360–363.
RYAN, J. J., PAOLO, A. M., & BRUNGARDT, T. M. (1990). Standardization of the Wechsler Adult Intelligence Scale — Revised for persons 75 years and older. Psychological Assessment, 2, 404–411.
SAAL, F. E., DOWNEY, R. G., & LAHEY, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. Psychological Bulletin, 88, 413–428.
SACCHI, C., & RICHAUD DE MINZI, M. C. (1989). The Holtzman Inkblot Technique in preadolescent personality. British Journal of Projective Psychology, 34 (2), 2–11.
SACKETT, P. R. (1994). Integrity testing for personnel selection. Current Directions in Psychological Science, 3, 73–76.
SACKETT, P. R., & WILK, S. L. (1994). Within-group norming and other forms of score adjustment in preemployment testing. American Psychologist, 49, 929–954.
SACKS, E. L. (1952). Intelligence scores as a function of experimentally established social relationships between the child and examiner. Journal of Abnormal and Social Psychology, 47, 354–358.
SADACCA, R., CAMPBELL, J. P., DIFAZIO, A. S., SCHULTZ, S. R., & WHITE, L. A. (1990). Scaling performance utility to enhance selection/classification decisions. Personnel Psychology, 43, 367–378.
SAKLOFSKE, D. H., & ZEIDNER, M. (Eds.). (1995). International handbook of personality and intelligence. New York: Plenum Press.
SALOVEY, P., & MAYER, J. D. (1990). Emotional intelligence. Imagination, Cognition, and Personality, 9, 185–211.
SALOVEY, P., & SLUYTER, D. J. (Eds.). (1997). Emotional development and emotional intelligence: Educational implications. New York: Basic Books.
SAMEJIMA, E (1969). Estimation of latent ability using a response pattern of graded scores. Psychometric Monograph, No. 17.
SAMUDA, R. J. (1975). Psychological testing of American minorities: Issues and consequences. New York: Dodd, Mead.
SAMUDA, R. J., KONG, S. L., CUMMINS, J., LEWIS, J., & PASCUAL-LEONE, J. (1991). Assessment and placement of minority students. Kirkland, WA: Hogrefe & Huber Publishers.
SANDOVAL, J. H., & MIILLE, M. P. W. (1980). Accuracy judgments of WISC-R item difficulty for minority groups. Journal of Consulting and Clinical Psychology, 48, 249–253.
SARASON, I. G. (1961). Test anxiety and the intellectual performance of college students. Journal of Educational Psychology, 52, 201–206.
SARASON, I. G. (Ed.). (1980). Test anxiety: Theory, research, and applications. Hillsdale, NJ: Erlbaum.

SARASON, S. B. (1954). The clinical interaction, with special reference to the Rorschach, New York: Harper.
SARASON, S. B., DAVIDSON, K. S., LIGHTHALL, F. F., WAITE, R. R., & RUEBUSH, B. K. (1960). Anxiety in elementary school children. New York: Wiley.
SARASON, S. B., HILL, K. T, & ZIMBARDO, P. (1964). A longitudinal study of the relation of test anxiety to performance on intelligence and achievement tests. Monographs of the Society for Research in Child Development, 29 (7, Serial No. 98).
SATTLER, J. M. (1970). Racial «experimenter effects» in experimentation, testing, and interviewing. Psychological Bulletin, 73, 137–160.
SATTLER, J. M. (1982). Assessment of children's intelligence and special abilities (2nd ed.). Boston: Allyn & Bacon.
SATTLER, J. M. (1988). Assessment of children (3rd ed.). San Diego, CA: Author.
SATTLER, J. M. (1992). Assessment of children: WISC-III and WPPSI-R supplement. San Diego, CA: Author.
SATTLER, J. M., & THEYE, F. (1967). Procedural, situational, and interpersonal variables in individual intelligence testing. Psychological Bulletin, 68, 347–360.
SAUDARGAS, R. A. (1989). Review of the Classroom Environment Scale, Second Edition. Tenth Mental Measurements Yearbook, 173–174.
SAUNDERS, D. R. (1956). Moderator variables in prediction. Educational and Psychological Measurement, 16, 209–222.
SAVICKAS, M. L., & LENT, R. W. (Eds.). (1994). Convergence in career development theories: Implications for science and practice. Palo Alto, CA: CPP Books.
SAX, G. (1991). The Fields Teaching Tests. Seattle: University of Washington.
SAXE, L. (1994). Detection of deception: Polygraph and integrity tests. Current Directions in Psychological Science, 3, 69–73.
SCARPATI, S. (1991). Current perspectives in the assessment of the handicapped. In R. K. Hambleton & J. N. Zaal (Eds.), Advances in educational and psychological testing: Theory and applications (pp. 251–276). Boston: Kluwer.
SCHAEFER, C. E., GITLIN, K., & SANDGRUND, A. (1991). Play diagnosis and assessment. New York: Wiley.
SCHAFER, W. O. (1992). Review of the Computer Programmer Aptitude Battery. Eleventh Mental Measurements Yearbook, 227–228.
SCHAIE, J. P. (1978). Review of the Gerontological Apperception Test. Eighth Mental Measurements Yearbook, Vol.1, 829–830.
SCHAIE, K. W. (1965V. A general model for the study of developmental problems. Psychological Bulletin, 64, 92–107.
SCHAIE, K. W. (1973). Methodological problems in descriptive developmental research on adulthood and aging. In J. R. Nesselroade & H. W. Reese (Eds.), Life-span developmental psychology: Methodological issues (pp. 253–280). New York: Academic Press.
SCHAIE, K. W. (1978). Review of the Senior Apperception Technique. Eighth Mental Measurements Yearbook, Vol. 1, 1060.
SCHAIE, K. W. (1988a). Internal validity threats in studies of adult cognitive development. In M. L. Howe & C. J. Brainard (Eds.), Cognitive development in adulthood: Progress in cognitive development research (pp. 241–272). New York: Springer-Verlag.
SCHAIE, K. W. (1988b). Manual for the Schaie-Thurstone Adult Mental Abilities Test (STAMAT). Palo Alto, CA: Consulting Psychologists Press.
SCHAIE, K. W. (1994). The course of adult intellectual development. American Psychologist, 49, 304–313.
SCHAIE, K. W., & GRIBBIN, K. (1975). Adult development and aging. Annual Review of Psychology, 26, 65–96.
SCHAIE, K. W., & HERTZOG, C. (1986). Toward a comprehensive model of adult intellectual development: Contributions of the Seattle Longitudinal Study. In R. J. Sternberg (Ed.), Advances in the psychology of human intelligence (Vol. 3, pp. 79–118). Hillsdale, NJ: Erlbaum.
SCHATZ, J., & HAMDAN-ALLEN, G. (1995). Effects of age and IQ on adaptive behavior domains for children with autism. Journal of Autism and Developmental Disorders, 25, 51–60.
SCHERICH, H.H.,&HANNA,G. S. (1977). Passage-dependence data in the selection of reading comprehension test items. Educational and Psychological Measurement, 37, 991–997.
SCHEUNEMAN, J. D. (1982). A posteriori analyses of biased items. In R. A. Berk (Ed.), Handbook of methods for detecting test bias (pp. 180–198). Baltimore: Johns Hopkins University Press.
SCHEUNEMAN, J., GERRITZ, K., & EMBRETSON, S. (1991). Effects of prose complexity on achievement test item difficulty (Res. Rep. No. 91–43). Princeton, NJ: Educational Testing Service.
SCHMID, J., & LEIMAN, J. (1957). The development of hierarchical factor solutions. Psychometrika, 22, 53–61.
SCHMIDT, F. L. (1985). Review of Wonderlic Personnel Test. Ninth Mental Measurements Yearbook, Vol. 2, 1755–1757.
SCHMIDT, F. L. (1992). What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. American Psychologist, 47, 1173–1181.
SCHMIDT, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. Psychological Methods, 1, 115–129.
SCHMIDT, F. L., BERNER, J. G., & HUNTER, J. E. (1973). Racial differences in validity of employment tests: Reality or illusion? Journal of Applied Psychology, 58, 5–9.
SCHMIDT, F. L., GAST-ROSENBERG, L., & HUNTER, J. E. (1980). Validity generalization results for computer programmers. Journal of Applied Psychology, 65, 643–661.
SCHMIDT, F. L., & HUNTER, J. E. (1977). Development of a general solution to the problem of validity generalization. Journal of Applied Psychology, 62, 529–540.
SCHMIDT, F. L., & HUNTER, J. E. (1992). Development of a casual model of processes determining job performance. Current Directions in Psychological Science, 1, 89–92.
SCHMIDT, F. L., HUNTER, J. E., McKENZIE, R. C., & MULDROW, T. W. (1979). Impact of valid selection procedures on work-force productivity. Journal of Applied Psychology, 64, 609–626.
SCHMIDT, F. L., HUNTER, J. E., & OUTERBRIDGE, A. N. (1986). Impact of job experience and ability on job knowledge, work sample performance, and supervisory ratings on job performance. Journal of Applied Psychology, 71, 432–439.
SCHMIDT, F. L., HUNTER, J. E., & PEARLMAN, K. (1981). Task differences as moderators of aptitude test validity in selection: A red herring. Journal of Applied Psychology, 66, 166–185.

SCHMIDT, F. L., HUNTER, J. E., PEARLMAN, K., & HIRSH, H. R. (1985). Forty questions about validity generalization and meta-analysis. Personnel Psychology, 38, 697–798.
SCHMIDT, F. L., HUNTER, J. E., PEARLMAN, K., & SHANE, G. S. (1979). Further tests of the Schmidt-Hunter Bayesian validity generalization model. Personnel Psychology, 32, 257–281.
SCHMIDT, F. L., HUNTER, J. E., & URRY, V. W. (1976). Statistical power in criterion' related validation studies. Journal of Applied Psychology, 61, 473–485.
SCHMIDT, F. L., LAW, K., HUNTER, J. E., ROTHSTEIN, H. R., PEARLMAN, K., & McDANIEL, M. (1993). Refinements in validity generalization methods: Implications for the situational specificity hypothesis. Journal of Applied Psychology, 78, 3–12.
SCHMIDT, F. L., ONES, D. S., & HUNTER, J. E. (1992). Personnel selection. Annual Review of Psychology, 43, 627–670.
SCHMIDT, F. L., PEARLMAN, K., & HUNTER, J. E. (1980). The validity and fairness of employment and educational tests for Hispanic Americans: A review and analysis. Personnel Psychology, 33, 705–724.
SCHMITT, N. (1995). Review of the Differential Aptitude Tests, Fifth Edition. Twelfth Mental Measurements Yearbook, 304–305.
SCHMITT, N., BORMAN, W. C., et al. (Eds.). (1993). Personnel selection in organizations. San Francisco: Jossey-Bass.
SCHMITT, N., MELLON, P. M., & BYLENGA, C. (1978). Sex differences in validity for academic and employment criteria, and different types of predictors. Journal of Applied Psychology, 63,145–150.
SCHNEIDER, W., & WEINERT, F. E. (Eds.). (1990). Interaction among aptitudes, strategies, and knowledge in cognitive performance. New York: Springer-Verlag.
SCHOENFELDT, L. F. (1985). Review of Wonderlic Personnel Test. Ninth Mental Measurements Yearbook, Vol. 2, 1755–1758.
SCHOENFELDT, L. F., & MENDOZA, J. L. (1991). The use of the computer in the practice of industrial/organizational psychology. In T. B. Gutkin & S. L. Wise (Eds.), The computer and the decision-making process (pp. 155–176). Hillsdale. NJ: Erlbaum.
SCHOENFELDT, L. F., & MENDOZA, J. L. (1994). Developing and using factorially derived biographical scales. In G. S. Stokes, M. D. Mumford, & W. A. Owens (Eds.), Biodata handbook: Theory, research, and use of biographical information in selection and performance prediction (pp. 147–169). Palo Alto, CA: Consulting Psychologists Press.
SCHOENFELDT, L. F., SCHOENFELDT, B. B., ACKER, S. R., & PERLSON, M. R. (1976). Content validity revisited: Test development of a content-oriented test of industrial reading. Journal of Applied Psychology, 61, 581–588.
SCHOGGEN, P. (1989). Behavior settings: A revision and extension of Roger G. Barker's «Ecological psychology.» Stanford, CA: Stanford University Press.
SCHULER, H., FARR, J. L., & SMITH, M. (Eds.). (1993). Personnel selection and assessment: Individual and organizational perspectives. Hillsdale, NJ: Erlbaum.
SCHULZ, R., & EWEN, R. B. (1993). Adult development and aging: Myths and emerging realities (2nd ed.) New York: Macmillan.
SCHWARTZ, M. M., COHEN, B. D., & PAVLIK, W. B. (1964). The effects of subject- and experimenter-induced defensive response sets on Picture-Frustration Test reactions. Journal of Projective Techniques, 28, 341–345.
SCHWARTZ, R. H. (1992). Is Holland's theory worthy of so much attention, or should vocational psychology move on? Journal of Vocational Behavior, 40, 179–187.
SCHWARTZ, S. H. (1992). Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. Advances in Experimental Social Psychology, 25, 1–65.
SCHWARTZ, S. H. (1994). Are there universal aspects in the structure and contents of human values? Journal of Social Issues, 50 (4), 19–45.
SCHWARTZ, S. H., & SAGIV, L. (1995). Identifying culture-specifics in the content and structure of values. Journal of Cross-Cultural Psychology, 26, 92–116.
SCHWARZ, P. A., & KRUG, R. E. (1972). Ability testing in developing countries: A handbook of principles and techniques. New York: Praeger.
SCHWARZER, R. (Ed.). (1992). Self-efficacy: Thought control of action. Washington, DC: Hemisphere.
SCIENCE RESEARCH ASSOCIATES. (1990). CRT Skills Test: Examiners manual. Rosemont, IL: Author.
SCOTTISH COUNCIL FOR RESEARCH IN EDUCATION (1949). The trend of Scottish intelligence. London: University of London Press.
SCRUGGS, C. (1994). [Review of Work Keys Assessments]. In J. T. Kapes, M. M. Mastie, & E. A. Whitfield (Eds.), A counselor's guide to career assessment instruments (3rd ed., pp. 126–130). Alexandria, VA: National Career Development Association.
SEASHORE, H. G. (1962). Women are more predictable than men. Journal of Counseling Psychology, 9, 261–270.
SEASHORE, H. G., WESMAN, A. G., & DOPPELT, J. E. (1950). The standardization of the Wechsler Intelligence Scale for Children. Journal of Consulting Psychology, 14, 99–110.
SECUREST, L. (1963). Incremental validity: A recommendation. Educational and Psychological Measurement, 23, 153–158.
SEGALL, M. H. (1983). On the search for the independent variable in cross-cultural psychology. In S. H. Irvine & J. W. Berry (Eds.), Human assessment and cultural factors (pp. 127–137). New York: Plenum Press.
SEGALL, M. H., CAMPBELL, D. T, & HERSKOVITS, M. J. (1966). The influence of culture on visual perception. Indianapolis, IN: Bobbs-Merrill.
SEGUIN, E. (1907). Idiocy: Its treatment by the physiological method. New York: Columbia University, Bureau of Publications. Teachers College. (Original work published 1866)
The Seventh Mental Measurements Yearbook. (1972). Highland Park, NJ: Gryphon Press.
SEXTON, D., KELLEY, M. E, & SURBECK, E. (1990). Piagetian-based assessment. In A. F. Rotatori, R. A. Fox, D. Sexton, & J. Miller (Eds.), Comprehensive assessment in special education: Approaches, procedures, and concerns (pp. 54–88). Springfield, IL: Charles C Thomas.
SEXTON, M. E. (1987). The correlates of sensorimotor functioning in infancy. In I. C. Uhgiris & J. McV. Hunt (Eds.), Infant performance and experience: New findings with the ordinal scales (pp. 230–251). Champaign: University of Illinois Press.
SHAFFER, M. B. (1985). Review of the Children's Apperception Test. Ninth Mental Measurements Yearbook, Vol. 1, 316–311.

SHAH, C. P., & BOYDEN, M. F. H. (1991). Assessment of auditory functioning. In B. A. Bracken (Ed.), The psychoeducational assessment of preschool children (2nd ed., pp. 341–378). Boston: Allyn & Bacon.
SHANKWEILER, D., CRAIN, S., KATZ, L., FOWLER, A. E., LIBERMAN, A. M., BRADY, S. A., THORNTON, R., LUNDQUIST, E., DREYER, L., FLETCHER, J. M., STUEBING, K. K., SHAYWITZ, S. E., & SHAYWITZ, B. A. (1995). Cognitive profiles of reading-disabled children: Comparison of language skills in phonology, morphology, and syntax. Psychological Science, 6, 149–156.
SHAPIRA, Z., & DUNBAR, R. L. M. (1980). Testing Mintzberg's managerial roles classification using an in-basket simulation. Journal of Applied Psychology, 65, 87–95.
SHAPIRO, D. L. (1991). Forensic psychological assessment: An integrative approach. Boston: Allyn & Bacon.
SHARF, J. C. (1994). The impact of legal and equal employment opportunity issues on personal history inquiries. In G. S. Stokes, M. D. Mumford, & W. A. Owens (Eds.), Biodata handbook: Theory, research, and use of biographical information in selection and performance prediction (pp. 351–390). Palo Alto, CA: Consulting Psychologists Press.
SHARP, S. E. (1898–1899). Individual psychology: A study in psychological method. American Journal of Psychology, 10, 329–391.
SHAVELSON, R. J., & BOLUS, R. (1982). Self-concept: The interplay of theory and methods. Journal of Educational Psychology, 74, 3–17.
SHAVELSON, R. J., HUBNER, J. J., & STANTON, G. C. (1976). Self-concept: Validation of construct interpretations. Review of Educational Research, 46, 407–441.
SHAVELSON, R. J., & WEBB, N. M. (1991). Generalizability theory: A primer. Newbury Park, CA: Sage.
SHAW, M. E., & WRIGHT, J. M. (1967). Scales for the measurement of attitudes. New York: McGraw-Hill.
SHAW, S. R., SWERDLIK, M. E., & LAURENT, J. (1993). Review of the WISC-III. Journal of Psychoeducational Assessment (Monograph Series: Advances in Psychoeducational Assessment). Germantown, TN: Psychoeducational Corporation.
SHEA, S. C. (1988). Psychiatric interviewing: The art of understanding. Philadelphia: Saunders.
SHEDLER, J., MAYMAN, M., & MANIS, M. (1993). The illusion of mental health. American Psychologist, 48, 1117–1131.
SHEEHAN, E. P (1995). Review of the Work Environment Scale, Second Edition. Twelfth Mental Measurements Yearbook, 1122–1123.
SHEEHAN, K., & MISLEVY, R. J. (1989). Integrating cognitive and psychometric models to measure document literacy (Res. Rep. No. 89–51). Princeton, NJ: Educational Testing Service.
SHELDON, W., & STEVENS, S. S. (1970). The varieties of temperament: A psychology of constitutional differences. New York: Hafner. (Original work published 1942)
SHEPARD, J. W. (1989). Review of the Jackson Vocational Interest Survey. Tenth Mental Measurements Yearbook, 403–404.
SHEPARD, L. A. (1984). Setting performance standards. In R. A. Berk (Ed.), A guide to criterion-referenced test construction (pp. 169–198). Baltimore: Johns Hopkins University Press.
SHERMAN, S. W., & ROBINSON, N. M. (Eds.). (1982). Ability testing of handicapped people: Dilemma for government, science, and the public. Washington, DC: National Academy Press.
SHINN, M. R. (1989). Curriculum-based measurement: Assessing special children. New York: Guilford Press.
SHINN, M. R., & BAKER, S. K. (1996). The use of curriculum-based measurement with diverse learners. In L. A. Suzuki, P. J. Meller, & J. G. Ponterotto (Eds.), Handbook of multicultural assessment: Clinical, psychological, and educational applications (pp. 179–222). San Francisco: Jossey-Bass.
SHINN, M. R., ROSENFIELD, S., & KNUTSON, N. (1989). Curriculum-based assessment: A comparison of models. School Psychology Review, 18, 299–316.
SHOCK, N. W., GREULICH, R. C., ANDRES, R., ARENBERG, D., COSTA, P. T., JR., LAKATTA, E. G., & TOBIN, J. D. (1984). Normal human aging: The Baltimore Longitudinal Study of Aging. Washington, DC: U.S. Government Printing Office. (NIH Publication No. 84–2450)
SHORE, C. W., & MARION, R. (1972). Suitability of using common selection test standards for Negro and white airmen (AFHRL-TR–72–53). Lackland Air Force Base, TX: Air Force Human Resources Laboratory, Personnel Research Division.
SHORE, T. H., SHORE, L. M., & THORNTON, G. C., III. (1992). Construct validity of self- and peer evaluations of performance dimensions in an assessment center. Journal of Applied Psychology, 77, 42–54.
SHUMAN, D. W. (1990). Law and mental health professionals: Texas, Washington, DC: American Psychological Association.
SHUMAN, D. W. (1993). Law and mental health professionals: Texas supplement, Washington, DC: American Psychological Association.
SHURRAGER, H. C., & SHURRAGER, P. S. (1964). Haptic Intelligence Scale for adult blind. Chicago: Stoelting.
SHWEDER, R. A., & SULLIVAN, M. A. (1993). Cultural psychology: Who needs it? Annual Review of Psychology, 44, 497–523.
SHYE, S. (1988). Inductive and deductive reasoning: A structural reanalysis of ability tests. Journal of Applied Psychology, 73, 308–311.
SIGEL, I. E. (1963). How intelligence tests limit understanding of intelligence. Merrill-Palmer Quarterly, 9, 39–56.
SIGI: A computer-based System of Interactive Guidance and Information. (1974–1975). Princeton, NJ: Educational Testing Service.
SILVERMAN, I., & SHULMAN, A. D. (1970). A conceptual model of artifact in attitude change studies. Sociometry, 33, 97–107.
SILVERMAN, L. H. (1959). A Q-sort study of the validity of evaluations made from projective techniques. Psychological Monographs, 73 (7, Whole No. 477).
SILVERSTEIN, A. B. (1982a). Alternative multiple-group solutions for the WISC and the WISC-R. Journal of Clinical Psychology, 38, 166–168.
SILVERSTEIN, A. B. (1982b). Factor structure of the Wechsler Adult Intelligence Scale — Revised. Journal of Consulting and Clinical Psychology, 50, 661–664.
SILVERSTEIN, A. B. (1986). Nonstandard standard scores on the Vineland Adaptive Behavior Scales: A cautionary note. American Journal on Mental Deficiency, 91, 1–4.
SILVERSTEIN, A. B. (1989). Review of the Multidimensional Aptitude Battery. Tenth Mental Measurements Yearbook, 523–524.

SILVERSTEIN, A. B. (1990). Short forms of individual intelligence tests. Psychological Assessment, 2,3–11.
SIMON, H. A. (1976). Identifying basic abilities underlying intelligent performance of complex tasks. In L. B. Resnick (Ed.), The nature of intelligence (pp. 65–98). Hillsdale, NJ: Erlbaum.
SIMON, H. A. (1990). Invariants of human behavior. Annual Review of Psychology, 41, 1–19.
SIMON, H. A. (1994). Focus on attention: The linkage between cognition and emotion. In W. Spaulding (Ed.), Nebraska Symposium on Motivation: Vol. 41. Integrative views of motivation, cognition, and emotion (pp. 1–21). Lincoln: University of Nebraska Press.
SINES, J. O. (1985). Review of Roberts Apperception Test for Children. Ninth Mental Measurements Yearbook, Vol. 2, 1290–1291.
SINGER, J. A., & SALOVEY, P. (1993). The remembered self: Emotion and memory in personality. New York: Free Press.
SIRECI, S. G., THISSEN, D., & WAINER, H. (1991). On the reliability of testlet-based tests. Journal of Educational Measurement, 28, 237–247.
SIVAN, A. B. (1991). Benton Visual Retention Test Fifth Edition: Manual. San Antonio, TX: Psychological Corporation.
SKINNER, E. A. (1995). Perceived control, motivation, and coping. Thousand Oaks, CA: Sage.
SKOLNICK, A. (1966). Motivational imagery and behavior over twenty years. Journal of Consulting Psychology, 30, 463–478.
SLAVIK, S. (1991). Early memories as a guide to client movement through life. Canadian Journal of Counseling, 25, 331–337.
SLEEK, S. (1995, June). APA's national college to begin issuing credentials. APA Monitor, p. 24.
SMITH, C. P. (1992). Reliability issues. In C. P. Smith (Ed.), Motivation and personality: Handbook of thematic content analysis (pp. 126–139). New York: Cambridge University Press.
SMITH, C. P. (Ed.), (with Atkinson, J. W., McClelland, D. C., & Veroff, J.). (1992). Motivation and personality: Handbook of thematic content analysis. New York: Cambridge University Press.
SMITH, C. R. (1989). Review of the Classroom Environment Scale, Second Edition. Tenth Mental Measurements Yearbook, 174–177.
SMITH, G. (1991). Assessing family interaction by the collaborative drawing technique. In C. E. Schaefer, K. Gitlin, & A. Sandgrund (Eds.), Play diagnosis and assessment (pp. 599–607). New York: Wiley.
SMITH, J., HARRE, R., & VAN LANGENHOVE, L. (Eds.). (1995). Rethinking psychology. Thousand Oaks, CA: Sage.
SMITH, P. B., & BOND, M. H. (1993). Social psychology across cultures: Analysis and perspectives. London: Harvester Wheatsheaf.
SMITTLE, P. (1990). Assessment's next wave: The computerized placement tests. College Board Review, No. 156, 22–27.
SNIDER, J. G., & OSGOOD, C. E. (Eds.). (1969). Semantic differential technique: A source-book. Chicago: Aldine.
SNOW, J. H. (1992). Review of the Luria-Nebraska Neuropsychological Battery: Forms I and II. Eleventh Mental Measurements Yearbook, 484–486.
SNOW, R. E. (1989). Toward assessment of cognitive and conative structures in learning. Educational Researcher, 18 (9), 8–14.
SNOW, R. E. (1990). Progress and propaganda in learning assessment [Review of the book Dynamic assessment: An interactional approach to evaluating learning potential]. Contemporary Psychology, 35, 1134–1136.
SNOW, R. E. (1992). Aptitude theory: Yesterday, today, and tomorrow. Educational Psychologist, 27, 5–32.
SNOW, R. E. (1993). Construct validity and constructed response tests. In R. E. Bennett & W. C. Ward (Eds.), Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment (pp. 45–60). Hillsdale, NJ: Erlbaum.
SNOW, R. E., & LOHMAN, D. E (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 263–331). New York: American Council on Education/Macmillan.
SNYDER, C. R., & LARSON, G. R. (1972). A further look at student acceptance of general personality interpretations. Journal of Consulting and Clinical Psychology, 38, 384–388.
SOCIETY FOR INDUSTRIAL AND ORGANIZATIONAL PSYCHOLOGY. (1987). Principles for the validation and use of personnel selection procedures (3rd ed.). College Park, MD: Author.
SOMMER, R. (1894). Diagnostic der Geisteskrankheiten fur praktische Arzte und Studierende. Wien & Leipzig: Urban & Schwarzenberg.
SONDEREGGER, T. B. (Ed.). (1992). Nebraska Symposium on Motivation: Psychology and aging. Lincoln: University of Nebraska Press.
SONTAG, L. W., BAKER, C. T, & NELSON, V. L. (1958). Mental growth and personality development: A longitudinal study. Monographs of the Society for Research in Child Development, 23 (2, Serial No. 68).
SPANGLER, W. D. (1992). Validity of questionnaire and TAT measures of need for achievement: Meta-analyses. Psychological Bulletin, 112, 140–154.
SPARROW, S. S., BALLA, D. A., & CICCHETTI, D. V. (1984a). Vineland Adaptive Behavior Scales: Interview Edition Expanded Form Manual. Circle Pines, MN: American Guidance Service.
SPARROW, S. S., BALLA, D. A., & CICCHETTI, D. V. (1984b). Vineland Adaptive Behavior Scales: Interview Edition Survey Form Manual. Circle Pines, MN: American Guidance Service.
SPAULDING, W. D. (Ed.). (1994). Integrative views of motivation, cognition, and emotion. Lincoln, NE: University of Nebraska Press.
SPEARMAN, C. (1904). «General intelligence» objectively determined and measured. American Journal of Psychology, 15, 201–293.
SPEARMAN, C. (1927). The abilities of man. New York: Macmillan.
SPENGLER, P. M., & STROHMER, D. C. (1994). Clinical judgmental biases: The moderating roles of counselor cognitive complexity and counselor client preferences. Journal of Counseling Psychology, 41, 8–17.
SPIELBERGER, C. D. (Ed.). (1972). Anxiety: Current trends in theory and research (Vol. 2). Orlando, FL: Academic Press.
SPIELBERGER, C. D. (1985). Assessment of state and trait anxiety: Conceptual and methodological issues. Southern Psychologist, 2, 6–16.
SPIELBERGER, C. D. (1988). State-Trait Anger Expression Inventory manual. Odessa, FL: Psychological Assessment Resources.
SPIELBERGER, C. D. (1989). State-Trait Anxiety Inventory: A comprehensive bibliography. Palo Alto, CA: Consulting Psychologists Press.
SPIELBERGER, C. D., et al. (1980). Test Anxiety Inventory: Preliminary professional manual. Palo Alto, CA: Consulting Psychologists Press.

SPIELBERGER, C. D., et al. (1983). Manual for the State-Trait Anxiety Inventory (STAI, Form Y). Palo Alto, CA: Consulting Psychologists Press.
SPIELBERGER, C. D., ANTON, W. D., & BEDELL, J. (1976). The nature and treatment of test anxiety. In M. Zuckerman & C. D. Spielberger (Eds.), Emotions and anxiety: New concepts, methods, and applications (pp. 317–345). New York: LEA/Wiley.
SPIELBERGER, C., & DIAZ-GUERRERO, R. (Eds.). (1990). Cross-cultural anxiety (Vol. 4). Bristol, PA: Hemisphere.
SPIELBERGER, C. D., GONZALEZ, H. P., & FLETCHER, T. (1979). Test anxiety reduction, learning strategies, and academic performance. In H. F. O'Neil, Jr. & C. D. Spielberger (Eds.), Cognitive and affective learning strategies (pp. 111–131). New York: Academic Press.
SPIELBERGER, C. D., GONZALEZ, H. P., TAYLOR, C. J., ALGAZE, B., & ANTON, W. D. (1978). Examination stress and test anxiety. In C. D. Spielberger & I. G. Sarason (Eds.), Stress and anxiety (Vol. 5, pp. 167–191). New York: Hemisphere.
SPIELBERGER, C. D., JOHNSON, E. H., RUSSELL, S. F., CRANE, R. J., JACOBS, G. A., & WORDEN, T. J. (1985). The experience and expression of anger: Construction and validation of an anger expression scale. In M. A. Chesney & R. H. Rosenman (Eds.), Anger and hostility in cardiovascular and behavioral disorders (pp. 5–30). New York: McGraw-Hill/Hemisphere.
SPIELBERGER, C. D., & SYDEMAN, S. J. (1994). State-Trait Anxiety Inventory and State-Trait Anger Expression Inventory. In M. Maruish (Ed.), The use of psychological testing for treatment planning and outcome assessment (pp. 292–321). Hillsdale, NJ: Erlbaum.
SPITZ, H. H. (1986). The raising of intelligence: Selected history of attempts to raise retarded intelligence. Hillsdale, NJ: Erlbaum.
SPRANGER, E. (1928). Types of men (P. J. W. Pigors, Trans.). Halle: Niemeyer.
SPREEN, O., & STRAUSS, E. (1991). A compendium of neuropsychological tests: Administration, norms, and commentary. New York: Oxford University Press.
SPRUILL, J. (1991). A comparison of the Wechsler Adult Intelligence Scale-Revised with the Stanford-Binet (4th edition) for mentally retarded adults. Psychological Assessment, 3,133–135.
STAABS, G. VON (1991). The Scenotest (J. A. Smith, Trans.). Toronto: Hogrefe & Huber. (Original work published 1964)
STAMOULIS, D. T., & HAUENSTEIN, N. M. A. (1993). Rater training and rating accuracy: Training for dimensional accuracy versus training for ratee differentiation. Journal of Applied Psychology, 78, 994–1003.
STANLEY, J. C. (Ed.). (1972). Preschool programs for the disadvantaged: Five experimental approaches to early childhood education. Baltimore: Johns Hopkins University Press.
STANLEY, J. C. (Ed.). (1973). Compensatory education for children, ages two to eight. Baltimore: Johns Hopkins University Press.
STARR, B. J., & KATKIN, E. S. (1969). The clinician as aberrant actuary: Illusory correlation and the Incomplete Sentences Blank. Journal of Abnormal Psychology, 74, 670–675.
STEELE, C. (Chair). (1995, August). Defying the Bell Curve – Social factors that inhibit and facilitate academic performance of women and minorities. Symposium at the annual convention of the American Psychological Association, New York.
STEELE, C., SPENCER, S., & ARONSON, J. (1995, August) Inhibiting the expression of intelligence: The role of stereotype vulnerability. In C. Steele (Chair), Defying the Bell Curve (Symposium conducted at the annual convention of the American Psychological Association, New York).
STEPHENSON, W. (1953). The study of behavior: Q-technique and its methodology. Chicago: University of Chicago Press.
STERNBERG, R. J. (1977). Intelligence, information processing, and analogical reasoning: The componential analysis of human abilities. Hillsdale, NJ: Erlbaum.
STERNBERG, R. J. (1980). Representation and process in linear syllogistic reasoning. Journal of Experimental Psychology, General, 109, 119–159.
STERNBERG, R. J. (1981). Testing and cognitive psychology. American Psychologist, 36, 1001–1011.
STERNBERG, R. J. (Ed.). (1982–1989). Advances in the psychology of human intelligence (Vols. 1–5). Hillsdale, NJ: Erlbaum.
STERNBERG, R. J. (1984). What cognitive psychology can (and cannot) do for test development. In B. S. Plake (Ed.), Social and technical issues in testing: Implications for test construction and usage (pp. 39–60). Hillsdale, NJ: Erlbaum.
STERNBERG, R. J. (1985a). Beyond IQ: A triarchic theory of human intelligence. New York: Cambridge University Press.
STERNBERG, R. J. (Ed.). (1985b). Human abilities: An information-processing approach. New York: Freeman.
STERNBERG, R. J. (1986). Intelligence applied: Understanding and increasing your intellectual skills. San Diego, CA: Harcourt Brace Jovanovich.
STERNBERG, R. J. (1988). Mental self-government: A theory of intellectual styles and their development. Human Development, 31, 197–224.
STERNBERG, R. J. (1989). The triarchic mind: A new theory of human intelligence. New York: Penguin.
STERNBERG, R. J. (1990). Metaphors of mind: Conceptions of the nature of intelligence. New York: Cambridge University Press.
STERNBERG, R. J. (1993). Rocky's back again: A review of the WISC-III. Journal of Psychoeducational Assessment (Monograph Series: Advances in Psychoeducational Assessment). Germantown, TN: Psychoeducational Corporation.
STERNBERG, R. J. (1994a). The PRSVL model of person-context interaction in the study of human potential. In M. G. Rumsey, C. B. Walker, & J. H. Harris (Eds.), Personnel selection and classification (pp. 317–332). Hillsdale, NJ: Erlbaum.
STERNBERG, R. J. (1994b). Thinking styles: Theory and assessment at the interface between intelligence and personality. In R. J. Sternberg & P. Ruzgis (Eds.), Personality and intelligence (pp. 169–187). New York: Cambridge University Press.
STERNBERG, R. J., & DETTERMAN, D. K. (Eds.). (1979). Human intelligence: Perspectives on its theory and measurement. Norwood, NJ: Ablex.
STERNBERG, R. J., & DETTERMAN, D. K. (Eds.). (1986). What is intelligence? Contemporary viewpoints on its nature and definitions. Norwood, NJ: Ablex.
STERNBERG, R. J., & FRENSCH, P. A. (Eds.). (1991). Complex problem solving: Principles and mechanisms. Hillsdale, NJ: Erlbaum.
STERNBERG, R. J., & RUZGIS, P. (Eds.). (1994). Personality and intelligence. New York: Cambridge University Press.

STERNBERG, R.J., & WAGNER, R. K. (Eds.). (1986). Practical intelligence: Origins of competence in the everyday world. New York. Cambridge University Press.
STERNBERG, R.J., WAGNER, R. K., WILLIAMS, W. M., & HORVATH, J. A. (1995). Testing common sense. American Psychologist, 50, 912–927.
STERNBERG, R.J., & WEIL, E. M. (1980). An aptitude x strategy interaction in linear syllogistic reasoning. Journal of Educational Psychology, 72, 226–239.
STEVENS, J. H., JR., & BAKEMAN, R. (1985). A factor analytic study of the HOME scale for infants. Developmental Psychology, 21, 1196–1203.
STEVENS, M. J., & CAMPION, M. A. (1994). Teamwork-KSA Test: Examiner's manual. Rosemont, IL: SRA-McGraw-Hill/London House.
STICHT, T. G. (Ed.). (1975). Reading for working: A functional literacy anthology. Alexandria, VA: Human Resources Research Organization.
STOKES, G. S., MUMFORD, M. D., & OWENS, W. A. (Eds.). (1994). Biodata handbook: Theory, research, and use of biographical information in selection and performance prediction. Palo Alto, CA: Consulting Psychologists Press.
STOKOLS, D. (1995). The paradox of environmental psychology. American Psychologist, 50, 821–837.
STOKOLS, D., & ALTMAN, I. (Eds.). (1987). Handbook of environmental psychology (Vols. 1 & 2). New York: Wiley.
STOLOFF, M. L., & COUCH, J. V. (Eds.). (1992). Computer Use in psychology: A directory of software (3rd ed.). Washington, DC: American Psychological Association.
STONE, B. J., GRIDLEY, B. E., & GYURKE, J. S. (1991). Confirmatory factor analysis of the WPPSI-R at the extreme end of the age range. Journal of Psychoeducational Assessment, 9, 263–270.
STONE, E. R, & HOLLENBECK, J. R. (1989). Clarifying some controversial issues surrounding statistical procedures for detecting moderators: Empirical evidence and related matters. Journal of Applied Psychology, 74, 3–10.
STONER, G. (1995). Review of the Metropolitan Readiness Tests, Fifth Edition. Twelfth Mental Measurements Yearbook, 612–614.
STORANDT, M., & VANDENBOS, G. R. (Eds.). (1994). Neuropsychological assessment of dementia and depression in older adults: A clinician's guide. Washington, DC: American Psychological Association.
STOTT, L. H. & BALL, S. (1965). Infant and preschool mental tests: Review and evaluation. Monographs of the Society for Research in Child Development, 30 (3, Serial No. 101).
STRAUSS, A. A., & LEHTINEN, L. E. (1947). Psychopathology and education of the brain-injured child. New York: Grune & Stratton.
STREINER, D. L., & NORMAN, G. R. (1995). Health measurement scales: A practical guide to their development and use (2nd ed.). Oxford, England: Oxford University Press.
STRICKER, G., DAVIS-RUSSELL, E., BOURG, E., DURAN, E., HAMMOND, W. R., McHOLLAND, W. R., POLITE, K., & VAUGHN, B. E. (Eds.). (1990). Toward ethnic diversification in psychology education and training. Washington, DC: American Psychological Association.
STRICKER, L. J. (1966). Compulsivity as a moderator variable: A replication and extension. Journal of Applied Psychology, 50, 331–335.
STRICKER, L. J. (1969). «Test-wiseness» on personality scales. Journal of Applied Psychology Monograph, 53(3, Part 2).
STRICKER, L. J. (1982). Interpersonal Competence Instrument: Development and preliminary findings. Applied Psychological Measurement, 6, 69–81.
STRICKER, L. J. (1984) Test disclosure and retest performance on the SAT. Applied Psychological Measurement, 8, 81–87.
STRICKER, L. J. (1985). Measuring social status with occupational information: A simple method (Res. Rep. 85–18). Princeton, NJ: Educational Testing Service.
STRICKER, L. J., & ROCK, D. A. (1990). Interpersonal competence, social intelligence, and general ability. Personality and Individual Differences, 11, 833–839.
Structural Equation Modeling: A Multidisciplinary Journal. Vol. 1. (1994). Hillsdale, NJ: Erlbaum.
STRUNK, W., JR., & WHITE, E. B. (1979). The elements of style (3rd ed.). Boston: Allyn & Bacon.
STURGIS, E. T, & GRAMLING, S. (1988). Psychophysiological assessment. In A. S. Bellack & M. Hersen (Eds.), Behavioral assessment: A practical handbook (3rd ed., pp. 213–251). New York: Pergamon Press.
SUBKOVIAK, M. J. (1984). Estimating the reliability of mastery-nonmastery classifications. In R. A. Berk (Ed.), A guide to criterion-referenced test construction (pp. 267–291). Baltimore: Johns Hopkins University Press.
SUBOTNIK, R. E, & ARNOLD, K. D. (Eds.). (1994). Beyond Terman: Contemporary longitudinal studies of giftedness and talent, Norwood, NJ: Ablex.
SUGARMAN, S. (1987). Piaget's construction of the child's reality. New York: Cambridge University Press.
SULLIVAN, P. M., & BURLEY, S. K. (1990). Mental testing of the hearing-impaired child. In C. R. Reynolds & R. W. Kamphaus (Eds.), Handbook of psychological and educational assessment of children: Intelligence and achievement (pp. 761–788). New York: Guilford Press.
SULLIVAN, P. M., & SCHULTE, L. E. (1992). Factor analysis of WISC-R with deaf and hard-of-hearing children. Psychological Assessment 4, 537–540.
SULSKY, L. M., & BALZER, W. K. (1988). Meaning and measurement of performance rating accuracy: Some methodological and theoretical concerns. Journal of Applied Psychology, 73, 1–10.
SULSKY, L. M., & DAY, D. V. (1992). Frame-of-reference training and cognitive categorization: An empirical investigation of rater memory issues. Journal of Applied Psychology, 7 7, 501–510.
SULSKY, L. M., & DAY, D. V. (1994). Effects of frame-of-reference training on rater accuracy under alternative time delays. Journal of Applied Psychology, 79, 535–543.
SUPER, D. E. (1953). A theory of vocational development. American Psychologist, 8, 185–190.
SUPER, D. E. (1957). The psychology of careers: An introduction to vocational development. New York: Harper & Row.
SUPER, D. E. (1980). A life-span, life-space approach to career development. Journal of Vocational Behavior, 16, 282–298.

SUPER, D. E. (1985). Coming of age in Middletown: Careers in the making. American Psychologist, 40, 405–415.
SUPER, D. E. (1990). A life-span, life-space approach to career development. In D. Brown, L. Brooks, et al. (Eds.), Career choice and development: Applying contemporary theories to practice (2nd ed., pp. 197–261). San Francisco: Jossey-Bass.
SUPER, D. E., et al. (1970). Computer-assisted counseling. New York: Teachers College Press.
SUPER, D. E., & BOHN, M. J., JR. (1970). Occupational psychology. Belmont, CA: Wadsworth.
SUPER, D. E., CRITES, J. O., HUMMEL, R. C., MOSER, H. P., OVERSTREET, P. L., & WARNATH, C. (1957). Vocational development: A framework for research. New York: Teachers College Press.
SUPER, D. E., & OVERSTREET, P. L. (1960). The vocational maturity of ninth grade boys. New York: Teachers College Press.
SUPER, D. E., & SVERKO, B. (Eds.). (1995). Life roles, values, and careers: International findings of the Work Importance Study. San Francisco: Jossey-Bass.
SUZUKI, L. A., MELLER, P. J., & PONTEROTTO, J. G. (Eds.). (1996). Handbook of multi-cultural assessment: Clinical, psychological, and educational applications. San Francisco: Jossey-Bass.
SWANSON, H. L., & KEOGH, B. (Eds.). (1990). Learning disabilities: Theoretical and research issues. Hillsdale, NJ: Erlbaum.
SWANSON, J. L. (1992). The structure of vocational interests for African-American college students. Journal of Vocational Behavior, 40, 144–157.
SWARTZ, J. D. (1973). Gamble's review of the Holtzman Inkblot Technique: Corrections and clarifications. Psychological Bulletin, 79, 378–379.
SWARTZ, J. D. (1992). The HIT and the HIT 25: Comments and clarifications. Journal of Personality Assessment, 58, 432–433.
SWARTZ, J. D., & HOLTZMAN, W. H. (1963). Group method of administration for the Holtzman Inkblot Technique. Journal of Clinical Psychology, 19, 433–441.
SWEZEY, R. W., & PEARLSTEIN, R. B. (1975). Guidebook for developing criterion-referenced tests. Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
SWINTON, S. S., & POWERS, D. E. (1985). The impact of self-study on GRE test performance (Res. Rep. 85–12). Princeton, NJ: Educational Testing Service.
SYMONDS, R M. (1931). Diagnosing personality and conduct. New York: Century.
SZYMULA, G. (1990). Vocational assessment. In C. Schiro-Geist (Ed.), Vocational counseling for special populations (pp. 65–97). Springfield, IL: Charles C Thomas.
TABE (1994). Complete battery, Forms 7 & 8. Examiners manual. Monterey, CA: CTB/ McGraw-Hill.
TAIT, M., PADGETT, M. Y., & BALDWIN, T. T. (1989). Job and life satisfaction: A reevaluation of the strength of the relationship and gender effects as a function of the date of the study. Journal of Applied Psychology, 74, 502–507.
TALLENT, N. (1992). The practice of psychological assessment. Englewood Cliffs, NJ: Prentice Hall.
TALLENT, N. (1993). Psychological report writing (4th ed.). Englewood Cliffs, NJ: Prentice Hall.
TAYLOR, H. C., & RUSSELL, J. T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection. Discussion and tables. Journal of Applied Psychology, 23, 565–578.
TAYLOR, S. E. (1990). Health psychology: The science and the field. American Psychologist, 45, 40–50.
TCS/2 technical report. (1993). Test of Cognitive Skills. Monterey, CA: CTB Macmillan/ McGraw-Hill.
TEETER, P. A. (1985). Review of Adjective Check List. Ninth Mental Measurements Yearbook, Vol. 1, 50–52.
TEGLASI, H. (1993). Clinical use of story telling: Emphasizing the T.A.T. with children and adolescents. Boston: Allyn & Bacon.
TELLEGEN, A., & BEN-PORATH, Y. S. (1992). The new uniform T scores for the MMPI–2: Rationale, derivation, and appraisal. Psychological Assessment, 4, 145–155.
TELLEGEN, A., & BEN-PORATH, Y. S. (1993). Code-type comparability of the MMPI and MMPI–2: Analysis of recent findings and criticisms. Journal of Personality Assessment, 61, 489–500.
TELZROW, C. F. (1990). Does PASS pass the test? A critique of the Das-Naglieri Cognitive Assessment System. Journal of Psychoeducational Assessment, 8, 344–355.
TENOPYR, M. L. (1986). Needed directions for measurement in work settings. In B. S. Plake & J. C. Witt (Eds.), The future of testing (pp. 269–288). Hillsdale, NJ: Erlbaum.
TENOPYR, M. L. (1989). Review of the Kuder Occupational Interest Survey, Revised (Form DD). Tenth Mental Measurements Yearbook, 427–429.
TENOPYR, M. L. (1995, August). Measurement at the crossroads. Presidential address (Div. 5) presented at the annual convention of the American Psychological Association, New York.
The Tenth Mental Measurements Yearbook. (1989). Lincoln, NE: Buros Institute of Mental Measurements.
TERMAN, L. M. (1916). The measurement of intelligence. Boston: Houghton Mifflin.
TERMAN, L. M., et al. (1925). Genetic studies of genius: Vol. I. Mental and physical traits of a thousand gifted children. Stanford University, CA: Stanford University Press.
TERMAN, L. M., & MERRILL, M. A. (1937). Measuring intelligence. Boston: Houghton Mifflin.
TERMAN, L. M., & MERRILL, M. A. (1960). Stanford-Binet Intelligence Scale: Manual for the third revision, Form L-M. Boston: Houghton Mifflin.
TERMAN, L. M., & MERRILL, M. A. (1973). Stanford-Binet Intelligence Scale: 1972 norms edition. Boston: Houghton Mifflin.
Tests in print II. (1974). Lincoln, NE: Buros Institute of Mental Measurements.
Tests in print III. (1983). Lincoln, NE: Buros Institute of Mental Measurements.
Tests in print IV (Vols. 1–2). (1994). Lincoln, NE: Buros Institute of Mental Measurements.
TETT, R. P., JACKSON, D. N., & ROTHSTEIN, M. (1991). Personality measures as predictors of job performance: A meta-analytic review. Personnel Psychology, 44, 703–742.
THARINGER, D. J., & STARK, K. (1990). A qualitative versus quantitative approach to evaluating the Draw-A-Person and Kinetic Family Drawing: A study of mood- and anxiety-disorder children. Psychological Assessment, 2, 365–375.
THOMAS, H. (1970). Psychological assessment instruments for use with human infants. Merrill-Palmer Quarterly of Behavioral Development, 16, 179–223.
THOMPSON, A. S., & LINDEMAN, R. H. (1981). Career Development Inventory: Vol. 1. Users manual. Palo Alto, CA: Consulting Psychologists Press.

THOMPSON, A. S., & LINDEMAN, R. H. (1984). Career Development Inventory: Vol. 2. Technical manual. Palo Alto, CA: Consulting Psychologists Press.

THOMPSON, D. (1995). Review of the Kuder General Interest Survey, Form E. Twelfth Mental Measurements Yearbook, 545–546.

THOMSON, G. H. (1948). The factorial analysis of human ability. (3rd ed.). Boston: Houghton Mifflin.

THORNDIKE, R. L. (1933). The effect of interval between test and retest on the constancy of the IQ. Journal of Educational Psychology, 24, 543–549.

THORNDIKE, R. L. (1940). «Constancy» of the IQ. Psychological Bulletin, 37, 167–186.

THORNDIKE, R. L. (1963). The concepts of over- and under-achievement. New York: Teachers College Press.

THORNDIKE, R. L. (1977). Causation of Binet IQ decrements. Journal of Educational Measurement, 14, 197–202.

THORNDIKE, R. L., HAGEN, E. P., & SATTLER, J. M. (1986a). The Stanford-Binet Intelligence Scale: Fourth Edition, Guide for administering and scoring, Chicago: Riverside.

THORNDIKE, R. L., HAGEN, E. P., & SATTLER, J. M. (1986b). The Stanford-Bind Intelligence Scale: Fourth Edition, Technical manual. Chicago: Riverside.

THORNDIKE, R. M. (1990). Would the real factors of the Stanford-Binet Fourth Edition please come forward? Journal of Psychoeducational Assessment, 8, 412–435.

THORNTON, G. C., III, & BYHAM, W. C. (1982). Assessment centers and managerial performance. Orlando, FL: Academic Press.

THORNTON, G. C., III, & ZORICH, S. (1980). Training to improve observer accuracy. Journal of Applied Psychology, 65, 351–354.

THURSTONE, L. L. (1925). A method of scaling psychological and educational tests. Journal of Educational Psychology, 16, 433–451.

THURSTONE, L. L. (1938). Primary mental abilities. Psychometric Monographs, No. 1.

TJHURSTONE, L. L. (1944). A factorial study of perception. Psychometric Monographs, No. 4.

THURSTONE, L. L. (1947a). The calibration of test items. American Psychologist, 2, 103–104.

THURSTONE, L. L. (1947b). Multiple factor analysis. Chicago: University of Chicago Press.

THURSTONE, L. L. (1950). Some primary abilities in visual thinking (No. 59). Chicago:
University of Chicago, Psychometric Laboratory.

THURSTONE, L. L. (1959). The measurement of values. Chicago: University of Chicago Press.

THURSTONE, L. L., & CHAVE, E. J. (1929). The measurement of attitude, Chicago: University of Chicago Press.

THURSTONE, L. L., AND THURSTONE, T. G. (1941). Factorial studies of intelligence. Psychometric Monographs, No. 2.

TIEDEMAN, D. V. (1994). «The guide is where?» [Review of the book Computer-assisted career decision making: The guide in the machine]. Contemporary Psychology, 39, 87–88.

TIMMONS, L. A., LANYON, R. I., ALMER, E. R., & CURRANT, P. J. (1993). Development and validation of sentence completion test indices of malingering during examination for disability. American Journal of Forensic Psychology, 11 (3), 23–38.

TITTLE, C. K. (1982). Use of judgmental methods in item bias studies. In R. A. Berk (Ed.), Handbook of methods for detecting test bias (pp. 31–63). Baltimore: Johns Hopkins University Press.

TITTLE, C. K., & ZYTOWSKI, D. G. (Eds.). (1978). Sex-fair interest measurement: Research and implications. Washington, DC: National Institute of Education.

TOBEY, L. H., & BRUHN, A. R. (1992). Early memories and the criminally dangerous. Journal of Personality Assessment, 59, 137–152.

TOPPING, D. M., CROWELL, D. C., & KOBAYASHI, V. N. (Eds.). (1989). Thinking across cultures: The Third International Conference on Thinking. Hillsdale, NJ: Erlbaum.

TORDY, G. R., EYDE, L. D., PRIMOFF, E. S., & HARDT, R. H. (1976). Job analysis of the position of New York State trooper: An application of the Job Element Method. Albany: New York State Police.

TOUYZ, S., BYRNE, D., & GILANDAS, A. (Eds.). (1994). Neuropsychology in clinical practice. San Diego, CA: Academic Press.

TRACEY, T. J., & ROUNDS, J. B. (1993). Evaluating Holland's and Gati's vocational interest models: A structural meta-analysis. Psychological Bulletin, 113, 229–246.

TRACEY, T. J. G., & ROUNDS, J. (1996). The spherical representation of vocational interests. Journal of Vocational Behavior, 48, 3–41.

TRAUB, R. E. (1993). On the equivalence of the traits assessed by multiple-choice and constructed-response tests. In R. E. Bennett & W. C. Ward (Eds.), Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment (pp. 29–44). Hillsdale, NJ: Erlbaum.

TRAXLER, A. E., & HILKERT, R. N. (1942). Effect of type of desk on results of machine-scored tests. School and Society, 56, 277–296.

TRENT, T., & LAURENCE, J. H. (Eds.). (1993). Adaptability screening for the Armed Forces. Washington, DC: Office of Assistant Secretary of Defense.

TREVISAN, M. S., SAX, G., & MICHAEL, W. B. (1991). The effects of the number of options per item and student ability on test validity and reliability. Educational and Psychological Measurement, 51, 829–831.

TREVISAN, M. S., SAX, G., & MICHAEL, W. B. (1994). Estimating the optimum number of options per item using an incremental option paradigm. Educational and Psychological Measurement, 54, 86–91.

TRIANDIS, H. C., DUNNETTE, M. D., & HOUGH, L. (Eds.). (1994). Handbook of industrial and organizational psychology (2nd ed., Vol. 4). Palo Alto, CA: Consulting Psychologists Press.

TRICKETT, E. J., & MOOS, R. H. (1995). Classroom Environment Scale manual: Development, applications, research (3rd ed.). Palo Alto, CA: Consulting Psychologists Press.

TRIMBLE, J. E., LONNER, W. J., & BOUCHER, J. D. (1983). Stalking the wily emic: Alternatives to cross-cultural measurement. In S. H. Irvine & J. W. Berry (Eds.), Human assessment and cultural factors (pp. 259–273). New York: Plenum Press.

TRYON, G. S. (1980). The measurement and treatment of test anxiety. Review of Educational Research, 50, 343–372.

TRYON, R. C. (1935). A theory of psychological components — an alternative to «mathematical factors.» Psychological Review, 42, 425–454.

TRYON, W. W. (Ed.). (1985). Behavioral assessment in behavioral medicine. New York: Springer.
TRYON, W. W. (1991). Activity measurement in psychology and medicine. New York: Plenum Press.
TRYON, W. W. (1996). Confidence interval testing: An alternative to null hypothesis testing. Manuscript submitted for publication.
TSUDZUKI, A., HATA, Y., & KUZE, T. (1957). [A study of rapport between examiner and subject.] Japanese Journal of Psychology, 27, 22–28.
TUDDENHAM, R. D. (1948). Soldier intelligence in World Wars I and II. American Psychologist, 3, 54–56.
TUDDENHAM, R. D., BLUMENKRANTZ, J., & WILKIN, W. R. (1968). Age changes on AGCT: A longitudinal study of average adults. Journal of Consulting and Clinical Psychology, 32, 659–663.
TURCO, T. L. (1989). Review of the Bracken Basic Concept Scale. Tenth Mental Measurements Yearbook, 102–104.
TURNBULL, W. W. (1985). Student change, program change: Why the SAT scores kept falling (College Board Rep. 85–2). New York: College Entrance Examination Board.
The Twelfth Mental Measurements Yearbook. (1995). Lincoln: Buros Institute of Mental Measurements.
TYLER, B., & MILLER, K. (1986). The use of tests by psychologists: Report on a survey of BPS members. Bulletin of the British Psychological Society, 39, 405–410.
TZINER, A., RONEN, S., HACOHEN, D. (1993). A four-year validation study of an assessment center in a financial corporation. Journal of Organizational Behavior, 14, 225–237.
UGUROGLU, M. E., & WALBERG, H. J. (1979). Motivation and achievement: A quantitative synthesis. American Educational Research Journal, 16, 375–389.
Uniform guidelines on employee selection procedures. (1978). Federal Register, 43 (166), 38296–38309.
U. S. DEPARTMENT OF DEFENSE. (1982). Profile of American youth: 1980 nationwide administration of the Armed Services Vocational Aptitude Battery. Washington, DC: Office of the Assistant Secretary of Defense (Manpower, Reserve Affairs, and Logistics).
U. S. DEPARTMENT OF LABOR. (1970). Manual for the USES General Aptitude Test Battery, Section III; Development. Washington, DC: U. S. Government Printing Office.
U. S. DEPARTMENT OF LABOR. (1979). Manual for the USES General Aptitude Test Battery: Section II. Occupational aptitude pattern structure. Washington, DC: U.S. Government Printing Office.
U. S. DEPARTMENT OF LABOR. (1980). Manual for the USES General Aptitude Test Battery: Section II-A. Development of the occupational aptitude pattern structure. Washington, DC: U.S. Government Printing Office.
U. S. DEPARTMENT OF LABOR. (1983a). The dimensionality of the General Aptitude Test Battery (GATB) and the dominance of general factors over specific factors in the prediction of job performance (USES Test Res. Rep. No. 44). Washington, DC: U.S. Government Printing Office.
U. S. DEPARTMENT OF LABOR. (1983b). The economic benefits of personnel selection using ability tests (USES Test Res. Rep. No. 47). Washington, DC: U. S. Government Printing Office.
U. S. DEPARTMENT OF LABOR. (1983c). Overview of validity generalization (USES Test Res. Rep. No. 43). Washington, DC: U.S. Government Printing Office.
U. S. DEPARTMENT OF LABOR. (1983d). Test validation/or 12,000 jobs: An application of job classification and validity generalization analysis to the General Aptitude Test Battery (USES Test Res. Rep. No. 45). Washington, DC: U.S. Government Printing Office.
U. S. DEPARTMENT OF LABOR EMPLOYMENT AND TRAINING ADMINISTRATION. (1991). Dictionary of occupational titles (4th rev. ed.). Washington, DC: Author.
Update on the new GRE General Test. (1995, Summer). GRE Board Newsletter, 10, 2–3.
URBINA, S. (1995). Review of the Basic Personality Inventory. Twelfth Mental Measurements Yearbook, pp. 105–106.
URBINA, S. (1997). Study guide: Psychological testing, Seventh edition. Upper Saddle River, NJ: Prentice Hall.
UŽGIRIS, I. C., & HUNT, J. McV. (1975). Assessment in infancy: Ordinal Scales of Psychological Development. Urbana, IL: University of Illinois Press.
UŽGIRIS, I. C., & HUNT, J. MCV. (Eds.). (1987). Infant performance and experience: New findings with the ordinal scales. Champaign: University of Illinois Press.
VACC, N. A. (1992). Review of the Career Assessment Inventory, Second Edition (Vocational version). Eleventh Mental Measurements Yearbook, 150–151.
VAIDYA, S., & CHANSKY, N. (1980). Cognitive development and cognitive style in mathematics achievement. Journal of Educational Psychology, 72, 326–330.
VAILLANT, G. E., & MCCULLOUGH, L. (1987). The Washington University Sentence completion Test compared with other measures of adult ego development. American Journal of Psychiatry, 144, 1189–1194.
VALCIUKAS, J. A. (1995). Forensic neuropsychology: Conceptual foundations and clinical practice. New York: Haworth Press.
VALENCIA, R. R. (1990). Clinical assessment of young children with the McCarthy Scales of Children's Abilities. In C. R. Reynolds & R. W. Kamphaus (Eds.), Handbook of psychological and educational assessment of children: Intelligence and achievement (pp. 209–258). New York: Guilford Press.
VALENCIA, R. R., & LOPEZ, R. (1992). Assessment of racial and ethnic minority students: Problems and prospects. In M. Zeidner & R. Most (Eds.), Psychological testing: An inside view (pp. 399–439). Palo Alto, CA: Consulting Psychologists Press.
VALENCIA, R. R., & RANKIN, R. J. (1985). Evidence of content bias on the McCarthy Scales with Mexican-American children: Implications for test translation and nonbiased assessment. Journal of Educational Psychology, 77, 197–207.
VANCE, H. B. (Ed.). (1993). Best practices in assessment for school and clinical settings. Bran-don, VT: Clinical Psychology.
Van Der MADE-VAN BEKKUM, I. J. (1971). Dutch word association norms. Amsterdam: Swets & Zeitlinger.
Van Der PLOEG, R. D. (Ed.). (1994a). Clinicians guide to neuropsychological assessment. Hillsdale, NJ: Erlbaum.
Van Der PLOEG, R. D. (1994b). Estimating premorbid level of functioning. In R. D. Van der Ploeg (Ed.), Clinicians guide to neuropsychological assessment (pp. 43–68). Hillsdale, NJ: Erlbaum.
Van de VIJVER, E, & HAMBLETON, R. K. (1996). Translating tests: Some practical guidelines. European Psychologist, 1, 89–99.

Van GORP, W. G. (1992). Review of the Luria-Nebraska Neuropsychological Battery: Forms I and II. Eleventh Mental Measurements Yearbook, 486–488.
Van SOMEREN, M., BARNARD, Y., & SANDBERG, J. (1994). The think aloud method: A practical guide to modeling cognitive processes. San Diego, CA: Academic Press.
VAZQUEZ NUTALL, E., ROMERO, I., & KALESNIK, J. (Eds.). (1992). Assessing and screening preschoolers: Psychological and educational dimensions (pp. 43–54). Boston: Allyn & Bacon.
VERHOEVE, M. A. (1993). JVIS applications handbook: A users guide for the Jackson Vocational Interest Survey. Port Huron, MI: Sigma Assessment Systems.
VERNON, P. E. (1960). The structure of human abilities (Rev. ed.). London: Methuen.
VERNON, P. E. (1969). Intelligence and cultural environment. London: Methuen.
VIGLIONE, D. J., JR. (1985). Review of the Rosenzweig Picture-Frustration Study. Ninth Mental Measurements Yearbook, Vol. 2, 1295–1297.
VIGLIONE, D. J. (1989). Rorschach science and art. Journal of Personality Assessment, 53, 195–197.
VINCENT, K. R., & HARMAN, M. J. (1991). The Exner Rorschach: An analysis of its clinical validity. Journal of Clinical Psychology, 47, 596–599.
VINITSKY, M. (1973). A forty-year follow-up on the vocational interests of psychologists and their relationship to career development. American Psychologist, 28, 1000–1009.
VITZ, P. C. (1990). The use of stories in moral development: New psychological reasons for an old education method. American Psychologist, 45, 709–720.
WACHS, T. D., & SHEEHAN, R. (Eds.). (1988). Assessment of young developmentally disabled children. New York: Plenum Press.
WACHTER, K. W., & STRAF, M. L. (Eds.). (1990). The future of meta-analysis. New York: Russell Sage Foundation.
WAGNER, E. E. (1985). Review of the Rosenzweig Picture-Frustration Study. Ninth Mental Measurements Yearbook, Vol. 2, 1297–1298.
WAHLSTROM, M., & BOERSMAN, F. J. (1968). The influence of test-wiseness upon achievement. Educational and Psychological Measurement, 28, 413–420.
WAINER, H. (1993a). Measurement problems. Journal of Educational Measurement, 30 (1), 1–21.
WAINER, H. (1993b). Some practical considerations when converting a linearly administered test to an adaptive format. Educational Measurement: Issues and Practice, 12 (1), 15–20.
WAINER, H., DORANS, N. J., FLAUGHER, R., GREEN, B. F., JR., MISLEVY, R. J., STEINBERG, L., & THISSEN, D. (1990). Computerized adaptive testing: A primer. Hillsdale, NJ: Erlbaum.
WAINER, H., & KIELY, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. Journal of Educational Measurement, 24, 185–201.
WAINER, H., & LEWIS, C. (1990). Toward a psychometrics for testlets. Journal of Educational Measurement, 27, 1–14.
WAITE, R. R., SARASON, S. B., LIGHTHALL, F. F, & DAVIDSON, K. S. (1958). A study of anxiety and learning in children. Journal of Abnormal and Social Psychology, 57, 267–270.
WALD, A. (1947). Sequential analysis. New York: Wiley.
WALD, A. (1950). Statistical decision function. New York: Wiley.
WALKER, B. S., & SPENGLER, P. M. (1995). Clinical judgment of major depression in AIDS patients: The effects of clinician complexity and stereotyping. Professional Psychology: Research and Practice, 26, 269–273.
WALLACE, S. R. (1965). Criteria for what? American Psychologist, 20, 411–417.
WALLER, N. G., LYKKEN, D. T., & TELLEGEN, A. (1995), Occupational interests, leisure time interests, and personality: Three domains or one? Findings from the Minnesota Twin Registry. In D. Lubinski & R. V. Dawis (Eds.), Assessing individual differences in human behavior: New concepts, methods, and findings (pp. 233–259). Palo Alto, CA: Davies-Black.
WALLER, N. G., & WALDMAN, I. D. (1990). A reexamination of the WAIS-R factor structure. Psychological Assessment, 2, 139–144.
WALSH, W. B., & BETZ, N. E. (1995). Tests and assessment (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.
WALSH, W. B., & OSIPOW, S. H. (Eds.). (1993). Career counseling for women. Hillsdale, NJ: Erlbaum.
WANG, M. C., REYNOLDS, M. C., & WALBERG, H. J. (Eds.). (1991). Handbook of special education: Research and practice, Vol. 4: Emerging programs. Elmsford, NY: Pergamon Press.
WARD, W. C., KLINE, R. G., & FLAUGHER, J. (1986). College Board Computerized Placement Tests: Validation of an adaptive test of basic skills (ETS Res. Rep. 86–29). Princeton, NJ: Educational Testing Service.
WARNER, W. L., MEEKER, M., & EELLS, K. (1949). Social class in America: A manual of procedure for the measurement of social status. Chicago: Science Research Associates.
WASIK, B. H., & WASIK, J. L. (1971). Performance of culturally deprived children on Concept Assessment Kit-Conservation. Child Development, 42, 1586–1590.
WATKINS, C. E. (1991). What have surveys taught us about the teaching and practice of psychological assessment? Journal of Personality Assessment, 56, 426–437.
WATKINS, C E., CAMPBELL, V. L., & NIEBERDING, R. (1994). The practice of vocational assessment by counseling psychologists. Counseling Psychologist, 22, 115–128.
WATKINS, C. E., JR., CAMPBELL, V. L., NIEBERDING, R., & HALLMARK, R. (1995). Contemporary practice of psychological assessment by clinical psychologists. Professional Psychology: Research and Practice, 26, 54–60.
WATKINS, M. W., & McDERMOTT, P. A. (1991). Psychodiagnostic computing: From interpretive programs to expert systems. In T. B. Gutkin & S. L. Wise (Eds.), The computer and the decision-making process (pp. 11–42). Hillsdale, NJ: Erlbaum.
WATSON, S. (1992). Review of the Test of Nonverbal Intelligence, Second Edition. Eleventh Mental Measurements Yearbook, 970–972.
WEBB, E. J., CAMPBELL, D. T., SCHWARTZ, R. D., SECHREST, L., & GROVE, J. B. (1981). Nonreactive measures in the social sciences (2nd ed.). Boston: Houghton Mifflin.

WEBSTER, E. C. (1982). The employment interview: A social Judgment process. Schomberg, Canada: S. I. P. Publications.
WECHSLER, D. (1939). The measurement of adult intelligence. Baltimore: Williams & Wilkins.
WECHSLER, D. (1958). The measurement and appraisal of adult intelligence (4th ed.). Baltimore: Williams & Wilkins.
WECHSLER, D. (1981). WAIS-R manual: Wechsler Adult Intelligence Scale — Revised. San Antonio, TX: Psychological Corporation.
WECHSLER, D. (1989). WPPSI-R: Manual. San Antonio, TX: Psychological Corporation.
WECHSLER, D. (1991). WISC-III: Manual. San Antonio, TX: Psychological Corporation.
WEDDING, D., & FAUST, D. (1989). Clinical judgement and decision making in neuropsychology. Archifes of Clinical Neuropsychology, 4, 233–265.
WEEKLEY, J. A., FRANK, B., O'CONNOR, E. J., & PETERS, L. H. (1985). A comparison of three methods of estimating the standard deviation of performance in dollars. Journal of Applied Psychology, 70, 122–126.
WEINER, I. B. (1994a). Rorschach assessment. In M. E. Maruish (Ed.), The use of psychological testing for treatment planning and outcome assessment (pp. 249–278). Hillsdale, NJ: Erlbaum.
WEINER, I. B. (1994b). The Rorschach Inkblot Method (RIM) is not a test: Implications for theory and practice. Journal of Personality Assessment, 62, 498–504.
WEINER, I. B. (1995a). How to anticipate ethical and legal challenges in personality assessments. In J. N. Butcher (Ed.), Clinical personality assessment: Practical approaches (pp. 95–103). New York: Oxford University Press.
WEINER, I. B. (1995b). Methodological considerations in Rorschach research. Psychological Assessment, 7, 330–337.
WEINER, I. B., & HESS, A. K. (Eds.). (1987). Handbook of forensic psychology. New York: Wiley.
WEISS, D. J. (1974). Strategies of adaptive ability measurement (Res. Rep. 74–5). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.
WEISS, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. Applied Psychological Measurement, 6, 473–492.
WEISS, D. J. (Ed.). (1983). New horizons in testing: Latent trait test theory and computerized adaptive testing. Orlando, FL: Academic Press.
WEISS, D. J., & BETZ, N. E. (1973). Ability measurement: Conventional or adaptive? (Res. Rep. 73–1). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.
WEISS, D. J., & DAVISON, M. L. (1981). Test theory and methods. Annual Review of Psychology, 32,629–658.
WEISS, D. J., & VALE, C. D. (1987). Computerized adaptive testing for measuring abilities and other psychological variables. In J. N. Butcher (Ed.), Computerized psychological assessment (pp. 325–343). New York: Basic Books.
WEISS, D. S., ZILBERG, N. J., & GENEVRO, J. L. (1989). Psychometric properties of Loeviriger's Sentence Completion Test in an adult psychiatric outpatient sample. Journal of Personality Assessment, 53, 478–486.
WEISSENBERG, P., & GRUENFELD, L. W. (1966). Relationships among leadership dimensions and cognitive style. Journal of Applied Psychology, 50, 392–395.
WELSH, G. S. (1956). Factor dimensions A and R. In G. S. Welsh & W. G. Dahlstrom (Eds.), Basic readings on the MMPI in psychology and medicine (pp. 264–281). Minneapolis: University of Minnesota Press.
WELSH, G. S. (1975a). Adjective Check List descriptions of Freud and Jung. Journal of Personality Assessment, 39, 160–168.
WELSH, G. S. (1975b). Creativity and intelligence: A personality approach. Chapel Hill: University of North Carolina, Institute for Research in Social Science.
WELSH, J. R., JR., WATSON, T. W., & REE, M. J. (1990). Armed Services Vocational Aptitude Battery (ASVAB): Predicting military criteria from general and specific abilities (AFHRL-TR-90-63). Brooks AFB, TX: U.S. Air Force Human Resources Laboratory.
WERNER, E. E., HONZIK, M. P., & SMITH, R. S. (1968). Prediction of intelligence and achievement at ten years from twenty months pediatric and psychological examinations. Child Development, 39, 1063–1075.
WERNER, H., & STRAUSS, A. A. (1941). Pathology of figure-background relation in the child. Journal of Abnormal and Social Psychology, 36, 236–248.
WERNER, H., & STRAUSS, A. A. (1943). Impairment in thought processes of brain-injured children. American Journal of Mental Deficiency, 47, 291–295.
WESMAN, A. G. (1949). Effect of speed on item-test correlation coefficients. Educational and Psychological Measurement, 9, 51–57.
WESMAN, A. G. (1952). Faking personality test scores in a simulated employment situation. Journal of Applied Psychology, 36, 112–113.
WEST, R. (1991). Computing for psychologists: Statistical analysis using SPSS and MINITAB. Langhorn, PA: Gordon & Breach.
WESTEN, D. (1991). Clinical assessment of object relations using the TAT. Journal of Personality Assessment, 56, 56–74.
WESTEN, D., LOHR, N., SILK, K. R., GOLD, L., & KERBER, K. (1990). Object relations and social cognition in borderlines, major depressives, and normals: A thematic apperception analysis. Psychological Assessment, 2, 355–364.
WESTENBERG, P. M., & BLOCK, J. (1993). Ego development and individual differences in personality. Journal of Personality and Social Psychology, 65, 792–800.
WETZLER, S. (1990). The Millon Clinical Multiaxial Inventory (MCMI): A review. Journal of Personality Assessment, 55, 445–464.
WHIMBEY, A. (1975). Intelligence can be taught. New York: Dutton.
WHIMBEY, A. (1977). Teaching sequential thought: The cognitive-skills approach. Phi Delta Kappan, 59, 255–259.
WHIMBEY, A. (1980). Students can learn to be better problem solvers. Educational Leadership, 37, 560–565.
WHIMBEY, A. (1990). Thinking through math word problems: Strategies for intermediate elementary school students. Hillsdale, NJ: Erlbaum.
WHIMBEY, A., & DENENBERG, V. H. (1966). Programming life histories: Creating individual differences by the experimental control of early experiences. Multivariate Behavioral Research, 1, 279–286.
WHITE, B. L. (1978). Experience and environment: Major influences on the development of the young child (Vol. 2). Englewood Cliffs, NJ: Prentice Hall.

WHITE, P. A. (1990). Ideas about causation in philosophy and psychology. Psychological Bulletin, 108, 3–18.
WHITE, R. F. (Ed.). (1992). Clinical syndromes in adult neuropsychology: The practitioner's handbook, Amsterdam: Elsevier.
WHITEMAN, M. (1964). Intelligence and learning. Merrill-Palmer Quarterly, 10, 297–309.
WHITEN, A. (Ed.). (1991). Natural theories of mind: Evolution, development, and simulation of everyday mind-reading. Oxford, England: Basil Blackwell.
WHITING, B. B. (1976). The problem of the packaged variable. In K. Riegel & J. Meacham (Eds.), The developing individual in a changing world (Vol.1, pp. 303–309). The Hague: Mouton.
WHITWORTH, J. R., & SUTTON, D. L. (1993). WISC–III compilation: What to do now that you know the score, Novato, CA: Academic Therapy Publications.
WHYTE, W. F. (1991). Social theory for action: How individuals and organizations learn to change. Newbury Park, CA: Sage.
WICKES, T. A., JR. (1956). Examiner influence in a testing situation. Journal of Consulting Psychology, 20, 23–26.
WIGDOR, A. K. (1982). Psychological testing and the law of employment discrimination. In A. K. Wigdor & W. R. Garner (Eds.), Ability testing: Uses, consequences, and controversies (Pt. 2, pp. 39–69). Washington, DC: National Academy Press.
WIGDOR, A. K., & GARNER, W. R. (Eds.). (1982). Ability testing: Uses, consequences, and controversies (Pts. 1 & 2). Washington, DC: National Academy Press.
WIGDOR, A. K., & GREEN, B. E., JR. (1991a). Performance assessment for the workplace (Vol. 1). Washington, DC: National Academy Press.
WIGDOR, A. K., & GREEN, B. R., JR. (1991b). Performance assessment for the workplace: Vol. 2. Technical issues. Washington, DC: National Academy Press.
WIGDOR, A. K., & SACKETT, P. R. (1993). Employment testing and public policy: The case of the General Aptitude Test Battery. In H. Schuler, J. L. Farr, & M. Smith (Eds.), Personnel selection and assessment: Individual and organizational perspectives (pp. 183–204). Hillsdale, NJ: Erlbaum.
WIGGINS, G. P. (1993). Assessing student performance: Exploring the purpose and limits of testing. San Francisco: Jossey-Bass.
WIGGINS, J. S. (1959). Interrelationships among MMPI measures of dissimulation under standard and social desirability instructions. Journal of Consulting Psychology, 23, 419–427.
WIGGINS, J. S. (1962). Strategic, method, and stylistic variance in the MMPI. Psychological Bulletin, 59, 224–242.
WIGGINS, J. S. (1966). Social desirability estimation and «faking good» well. Educational and Psychological Measurement, 26, 329–341.
WIGGINS, J. S. (1988). Personality and prediction: Principles of personality assessment. Malabar, FL: S. A. Krieger. (Original work published 1973)
WIGGINS, J. S. (1989). Review of the Myers-Briggs Type Indicator. Tenth Mental Measurements Yearbook, 536–538.
WIGGINS, J. S. (1996). An informal history of the interpersonal circumplex tradition. Journal of Personality Assessment, 66, 217–233.
WIGGINS, J. S., & PINCUS, A. L. (1992). Personality: Structure and Assessment. Annual Review of Psychology, 43, 493–504.
WIGGINS, N. (1966). Individual viewpoints of social desirability. Psychological Bulletin, 66, 68–77.
WIIG, E. H. (1985). Review of Peabody Picture Vocabulary Test — Revised. Ninth Mental Measurements Yearbook, Vol. 2, 1127–1128.
WILLETT, J. B., & SAYER, A. G. (1994). Using covariance structure analysis to detect correlates and predictors of individual change over time. Psychological Bulletin, 116, 363–381.
WILLIAMS, A. (1995). Visual and active supervision: Roles, focus, technique. New York: Norton.
WILLIAMS, C. L., BUTCHER, J. N., BEN-PORATH, Y. S., & GRAHAM, J. R. (1992). MMPI-A Content scales: Assessing psychopathology in adolescents. Minneapolis: University of Minnesota Press.
WILLIAMS, H. G. (1991). Assessment of gross motor functioning. In B. A. Bracken (Ed.), The psychoeducational assessment of preschool children (2nd ed., pp. 284–316). Boston: Allyn & Bacon.
WILLIAMS, M. (1960). The effect of past experience on mental performance in the elderly. British Journal of Medical Psychology, 33, 215–219.
WILLINGHAM, W. W. (1988). Testing handicapped people: The validity issue. In H. Wainer & H. Braun (Eds.), Test validity (pp. 89–103). Hillsdale, NJ: Erlbaum.
WILLINGHAM, W. W., RAGOSTA, M., BENNETT, R. E., BRAUN, H., ROCK, D. A., & POWERS, D. E. (1988). Testing handicapped people. Boston: Allyn & Bacon.
WILLIS, J. (1970). Group versus individual intelligence tests in one sample of emotionally disturbed children. Psychological Reports, 27, 819–822.
WILLIS, S. L., BLIESZNER, R., & BALTES, P. B. (1981). Intellectual training research in aging: Modification of performance on the fluid ability of figural relations. Journal of Educational Psychology, 73, 41–50.
WILLIS, S. L., & SCHAIE, K. W. (1986). Practical intelligence in later adulthood. In R. J. Sternberg & R. K. Wagner (Eds.), Practical intelligence: Origins of competence in the everyday world (pp. 236–268). New York: Cambridge University Press.
WILLOCK, B. (1992). Projection, transitional phenomena, and the Rorschach. Journal of Personality Assessment, 59, 99–116.
WILLSON, V. L. (1994). Cognitive modeling of individual responses in test design. In C. R. Reynolds (Ed.), Cognitive assessment: A multidisciplinary perspective (pp. 155–173). New York: Plenum Press.
WILSON, R. S., & MATHENY, A. P., JR. (1983). Assessment of temperament in infant twins. Developmental Psychology, 19, 172–183.
WILSON, S. L. (1991). Microcomputer-based psychological assessment: An advance in helping severely physically disabled people. In P. L. Dann, S. H. Irvine, & J. M. Collis (Eds.), Advances in computer-based human assessment (pp. 171–187). Dordrecht, The Netherlands: Kluwer.
WINK, P. (1991). Two faces of narcissism. Journal of Personality and Social Psychology, 61, 590–597.
WINK, P. (1992). Three narcissism scales for the California Q-set. Journal of Personality Assessment, 58, 51–66.
WINK, P., & HELSON, R. (1993). Personality change in women and their partners. Journal of Personality and Social Psychology, 65, 597–605.

WINTER, D. A. (1992). Personal construct psychology in clinical practice: Theory, research, and applications. New York: Routledge, Chapman & Hall.
WIRT, R. D., & LACHAR, D. (1981). The Personality Inventory for Children: Development and clinical applications. In P. McReynolds (Ed.), Advances in psychological assessment (Vol. 5, pp. 353–392). San Francisco: Jossey-Bass.
WIRT, R. D., LACHAR, D., KLINEDINST, J. K., & SEAT, P. D. (1991). Multidimensional description of child personality: A manual for the Personality Inventory for Children 1990 Edition. Los Angeles: Western Psychological Services.
WIRTZ, W. (Chair). (1977). On further examination: Report of the Advisory Panel on the Scholastic Aptitude Test Score Decline. New York: College Entrance Examination Board.
WISE, L. L., MCHENRY, J., & CAMPBELL, J. P. (1990). Identifying optimal predictor composites and testing for generalizability across jobs and performance factors. Personnel Psychology, 43, 355–366.
WISE, P. S. (1989). The use of assessment techniques by applied psychologists. Belmont, CA: Wadsworth.
WISKOFF, M. R, & SCHRATZ, M. K. (1989). Computerized adaptive testing of a vocational aptitude battery. In R. F. Dillon & J. W. Pellegrino (Eds.), Testing: Theoretical and applied perspectives (pp. 66–96). New York: Praeger.
WISSLER, C. (1901). The correlation of mental and physical traits. Psychological Monographs, 3(6, Whole No. 16).
WITKIN, H. A., DYK, R. B., FATERSON, H. F., GOODENOUGH, D. R., & KARP, S. A. (1974). Psychological differentiation: Studies in development. New York: Wiley. (Original work published in 1962)
WITKIN, H. A., & GOODENOUGH, D. R. (1977). Field dependence and interpersonal behavior. Psychological Bulletin, 84, 661–689.
WITKIN, H. A., & GOODENOUGH, D. R. (1981). Cognitive styles: Essence and Origins — Field dependence and independence. New York: International Universities Press.
WITKIN, H. A., LEWIS, H.B., HERTZMAN, M., MACHOVER, K., MEISSNER, P. B., & WAPNER, S. (1972). Personality through perception: An experimental and clinical study. Westport, CT: Greenwood Press. (Original work published 1954)
WITKIN, H. A., OLTMAN, P. K., RASKIN, E., & KARP, S. A. (1971). A manual for the Embedded Figures Tests. Palo Alto, CA: Consulting Psychologists Press.
WITKIN, H. A., PRICE-WILLIAMS, D., BERTINI, M., CHRISTIANSEN, B., OLTMAN, P. K., RAMIREZ, M., & VAN MEEL, J. (1974). Social conformity and psychological differentiation. International Journal of Psychology, 9, 11–29.
WITT, J. C, ELLIOTT, S. N., GRESHAM, R M., & KRAMER, J. J. (1988). Assessment of special children: Tests and the problem-solving process. Glenview, IL: Scott, Foresman.
WITT, J. C., HEFFER, R. W., & PFEIFFER, J. (1990). Structured rating scales: A review of self-report and informant rating processes, procedures, and issues. In C. R. Reynolds & R. W. Kamphaus (Eds.), Handbook of psychological and educational assessment of children: Personality, behavior, and context (pp. 364–394). New York: Guilford Press.
WOLF, D. P. (1993). Assessment as an episode of learning. In R. E. Bennett & W. C. Ward (Eds.), Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment (pp. 213–240). Hillsdale, NJ: Erlbaum.
WOLF, E. M. (1986). Meta-analysis: Quantitative methods for research synthesis. Newbury Park, CA: Sage.
WOLF, T. H. (1973). Alfred Binet. Chicago: University of Chicago Press.
WOLK, R. L., & WOLK, R. B. (1971). Manual: Gerontological Apperception Test. New York: Human Sciences Press.
WOMER, F. B. (1970). What is National Assessment? Ann Arbor, MI: National Assessment of Educational Progress.
WOMER, M. (1972). Culture and the concept of intelligence: A case in Uganda. Journal of Cross-Cultural Psychology, 3, 327–328.
Wonderlic Personnel Test, Inc. (1992). Wonderlic Personnel Test & Scholastic Level Exam: User's manual. Libertyville, IL: Author.
WOOD, J. M., NEZWORSKI, M. T., & STEJSKAL, W. J. (1996a). The Comprehensive System for the Rorschach: A critical examination. Psychological Science, 7, 3–10.
WOOD, J. M., NEZWORSKI, M. T., & STEJSKAL, W. J. (1996b). Thinking critically about the Comprehensive System for the Rorschach: A reply to Exner. Psychological Science, 7, 14–17.
WOODCOCK, R. W., & JOHNSON, M. B. (1989, 1990). Woodcock-Johnson Psycho-Educational Battery — Revised. Alien, TX: DLM Teaching Resources.
WOOTEN, K. C., BARNER, B. O., & SILVER, N. C. (1994). The influence of cognitive style upon work environment preferences. Perceptual and Motor Skills, 79, 307–314.
WORCHEL, F. F., & DUPREE, J. L. (1990). Projective storytelling techniques. In C. R. Reynolds & R. W. Kamphaus (Eds.), Handbook of psychological and educational assessment of children: Personality, behavior, and context (pp. 70–88). New York: Guilford Press.
WORTHEN, B. R. (1995). Review of the Strong Interest Inventory (Fourth Edition). Twelfth Mental Measurements Yearbook, 999–1002.
WRIGHT, B. D. (1977). Solving measurement problems with the Rasch model. Journal of Educational Measurement, 14, 97–116.
WRIGHT, B. D., & STONE, M. H. (1979). Best test design: Rasch measurement. Chicago: Mesa Press.
WULACH, J. S. (1991). Law and mental health professionals: New Jersey. Washington, DC: American Psychological Association.
WYLIE, R. C. (1989). Measures of self-concept. Lincoln: University of Nebraska Press.
YAMA, M. F. (1990). The usefulness of human figure drawings as an index of overall adjustment. Journal of Personality Assessment, 54, 78–86.
YARROW, L.J., MACTURK, R. H., VIETZE, P. M., McCARTHY, M. E., KLEIN, R. P., & McQUISTON, S. (1984). Developmental course of parental stimulation and its relationship to mastery motivation during infancy. Developmental Psychology, 20, 492–503.
YARROW, L.J., McQUISTON, S., MACTURK, R. H., McCARTHY, M. E., KLEIN, R., & VIETZE, P. M. (1983). Assessment of mastery motivation during the first year of life: Contemporaneous and cross-age relationships. Developmental Psychology, 19, 159–171.
YARROW, L.J., & MESSER, D.J. (1983). Motivation and cognition in infancy. In M. Lewis (Ed.), Origins of intelligence: Infancy and early childhood (2nd ed., pp. 451–477). New York: Plenum Press.

YARROW, L. J., & PEDERSEN, F. A. (1976). The interplay between cognition and motivation in infancy. In M. Lewis (Ed.), Origins of intelligence: Infancy and early childhood (pp. 379-399). New York: Plenum Press.
YATES, A. J. et al. (1953-1954). Symposium on the effects of coaching and practice in intelligence tests. British Journal of Educational Psychology, 23, 147-162; 24, 1-8, 57-63.
YERKES, R. M. (Ed.). (1921). Psychological examining in the United States Army. Memoirs of the National Academy of Sciences, Vol. 15.
YORK, K. L., & JOHN, O. P. (1992). The four faces of Eve: A typological analysis of women's personality at midlife. Journal of Personality and Social Psychology, 63, 494-508.
YOUNG, F. W. (1984). Scaling. Annual Review of Psychology, 35, 55-81.
YSSELDYKE, J. E. (1989). Review of the Bracken Basic Concept Scale. Tenth Mental Measurements Yearbook, 104-105.
YUKL, G., & VAN FLEET, D. D. (1992). Theory and research on leadership in organizations. In M. D. Dunnette & L. M. Hough (Eds.), Handbook of industrial and organizational psychology (2nd ed., Vol. 3, pp. 147-197). Palo Alto, CA: Consulting Psychologists Press.
ZACHARY, R. A. (1990). Wechsler's Intelligence Scales: Theoretical and practical considerations. Journal of Psychoeducational Assessment, 8, 276-289.
ZARSKE, J. A. (1985). Review of Adjective Check List. Ninth Mental Measurements Yearbook, Vol 1, 52-53.
ZEDECK, S. (1971). Problems with the use of «moderator» variables. Psychological Bulletin, 76,295-310.
ZEICHMEISTER, E. B., & JOHNSON, J. E. (1992). Critical thinking: A functional approach. Pacific Grove, CA: Brooks/Cole.
ZEIDNER, J., & JOHNSON, C. D. (1991). Classification efficiency and systems design. Journal of the Washington Academy of Sciences, 81, 110-128.
ZEIDNER, M. (1987). Test of the cultural bias hypothesis: Some Israeli findings. Journal of Applied Psychology, 72, 38-48.
ZEIDNER, M. (1988). Cultural fairness in aptitude testing revisited: A cross-cultural parallel. Professional Psychology, Research and Practice, 19, 257-262.
ZEIDNER, M. (1993). Essay versus multiple-choice type classroom exams: The student's perspective. In B. Nevo & R. S. Jager (Eds.), Educational and psychological testing: The test taker's outlook (pp. 67-82). Toronto, Canada: Hogrefe & Huber.
ZEIDNER, M. (1995). Personality trait correlates of intelligence. In D. H. Saklofske & M. Zeidner (Eds.), International handbook of personality and intelligence (pp. 299-319). New York: Plenum Press.
ZELNIKER, T. (1989). Cognitive style and dimensions of information processing. In T. Globerson & T. Zeiniker (Eds.), Cognitive style and cognitive development (pp. 172-191). Norwood, NJ: Ablex.
ZENDERLAND, L. (1987). The debate over diagnosis: Henry Hebert Goddard and the medical acceptance of intelligence testing. In M. M. Sokal (Ed.), Psychological testing and American society, 1890-1930 (pp. 46-74). New Brunswick, NJ: Rutgers University Press.
ZIGLER, E., & MUENCHOW, S. (1992). Head Start: The inside story of America's most successful educational experiment. New York: Basic Books. ZIGLER, E., & STYFCO, S. J. (Eds.). (1993). Head Start and beyond. New Haven, CT: Yale University Press.
ZIGLER, E., & VALENTINE, J. (Eds.). (1980). Project Head Start: A legacy of the war on poverty. New York: Free Press.
ZIMMERMAN, B. J., & ROSENTHAL, T. L. (1974a). Conserving and retaining equalities and inequalities through observation and correction. Developmental Psychology, 10, 260-268.
ZIMMERMAN, B. J., & ROSENTHAL, T. L. (1974b). Observational learning of rule-governed behavior by children. Psychological Bulletin, 81, 29-42.
ZIMMERMAN, I. L., & WOO-SAM, J. (1972). Research with the Wechsler Intelligence Scale for Children: 1960-1970 [Special Monograph Suppl.] Psychology in the Schools, 9, 232-271.
ZUCKERMAN, M., KUHLMAN, D. M., JOIREMAN, J., TETA, P., & KRAFT, M. (1993). A comparison of three structural models for personality: The big three, the big five, and the alternative five. Journal of Personality and Social Psychology, 65, 757-768.
ZYTOWSKI, D. G. (1992). Three generations: The continuing evolution of Frederic Kuder's interest inventories. Journal of Counseling and Development, 71, 245-248.
ZYTOWSKI, D. G., & BORGEN, F. H. (1983). Assessment. In B. Walsh & S. H. Osipow (Eds.), Handbook of vocational psychology: Vol. 2. Applications (pp. 5-45). Hillsdale, NJ: Erlbaum.
ZYTOWSKI, D. G., & WARMAN, R. E. (1982). The changing use of tests in counseling. Measurement and Evaluation in Guidance, 15, 147-152.

АЛФАВИТНО-ПРЕДМЕТНЫЙ УКАЗАТЕЛЬ

Анастази Анна, Урбина Сьюзан

Психологическое тестирование

7-е издание

Перевод с английского и общая научная редакция профессора А. А. Алексеева

Заведующий редакцией	*П. Алесов*
Художественный редактор	*Е. Дьяченко*
Литературные редакторы	*С. Комаров, В. Попов*
Корректор	*М. Рошаль*
Верстка	*А. Рапопорт*

Подписано в печать 31.01.07. Формат 70×100/16. Усл. п. л. 54,18. Доп. тираж 3000 экз. Заказ № 3935.

ООО «Питер Пресс». Санкт-Петербург, Петергофское шоссе, д. 73, лит А29.

Налоговая льгота — общероссийский классификатор продукции ОК 005-93, том 2;
953005 — литература учебная.

Отпечатаио с фотоформ в ОАО «Печатный двор» им. А. М. Горького.
197110, Санкт-Петербург, Чкаловский пр., 15.

Классическая работа Анны Анастази «Психологическое тестирование» по праву считается «энциклопедией западной тестологии». При подготовке 7-го издания, выпущенного в США в 1997 году, текст книги был основательно переработан. Появилось несколько новых глав, написанных соавтором А. Анастази — С. Урбина. Содержательные изменения отражают новейшие тенденции развития психологического тестирования, в том числе возрастающее влияние компьютеризации как фактора интеграции психологической науки в целом и методов тестирования в частности. В новом издании уделено значительное внимание компьютеризированному адаптивному тестированию, метаанализу, моделированию структурными уравнениями, использованию доверительных интервалов, кросс-культурному тестированию, применению факторного анализа в разработке тестов личности и способностей и другим широко используемым и быстро развивающимся понятиям и процедурам, которые будут оказывать влияние на психометрическую практику в XXI веке.



Заказ книг:

197198, Санкт-Петербург, а/я 619
тел.: (812) 703-73-74, postbook@piter.com

61093, Харьков-93, а/я 9130
тел.: (057) 712-27-05, piter@kharkov.piter.com

www.piter.com — вся информация о книгах и веб-магазин

ISBN 978-5-272-00106-1

9 785272 001061